GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master's Thesis

submitted in partial fulfillment of the
requirements for the course "Applied Computer Science"

Navigating the Past: Using Annotations and LLMs for Corpus Preparation and Identifying Valid Itineraries in Historical Travelogues

Tarjia Alam Nisha

Institute of Computer Science

Master's Thesis
of the Center for Computational Sciences
at the Georg-August-Universität Göttingen

03. March 2025

Georg-August-Universität Göttingen
Institute of Computer Science

Goldschmidtstraße 7
37077 Göttingen
Germany

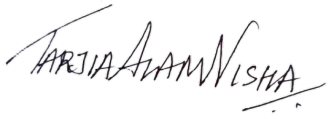
☎ +49 (551) 39-172000
FAX +49 (551) 39-14403
✉ office@informatik.uni-goettingen.de
www.informatik.uni-goettingen.de

First Supervisor: Dr. Joerg Wettlaufer
Second Supervisor: Prof. Dr. Caroline Sporleder
Third Supervisor: Dr. Franziska Pannach

I hereby declare that I have written this thesis independently without any help from others and without the use of documents or aids other than those stated. I have mentioned all used sources and cited them correctly according to established academic citation rules.

In this paper, I have used ChatGPT for proofreading and optimizing or restructuring the texts. I hereby declare that I have stated all uses completely. Missing or incorrect information will be considered as an attempt to cheat.

Göttingen, 03. March 2025



Abstract

In historical travelogues, integrating the mapping of place and person entities while accounting for their travel connections enhances the understanding of itinerary structures. Previous methods have predominantly analyzed place and person entities independently, utilizing essential Named Entity Recognition (NER) techniques or associating persons with places through spatial categories within a specified narrative timeframe. However, these approaches struggle to identify valid itineraries, as texts often refer to places that are not part of the journey. To address this limitation, this study introduces a fundamental methodology that categorizes places based on their travel relationships, distinguishing whether they are part of the journey or not. The primary objective of this research is to accurately extract and validate itineraries from English historical travelogues by assigning travel-related categories—such as valid connections between characters, character-specific travel connections, and identification of places integral to the journey—to co-occurring person and place entities within text segments. Additionally, this categorization considers the linguistic context, spatial relationships, and temporal dimensions. This approach aims to systematically reconstruct and interpret travel narratives by developing comprehensive annotation guidelines, ensuring a thorough understanding of historical journeys and their contextual significance. These guidelines will facilitate accurately identifying and categorizing travel-related elements, such as character interactions, travel connections, and journey-specific locations, enabling a structured and reliable analysis of historical travelogues. For evaluation purposes, a corpus of 120 pages drawn from 19th-century English-language travelogues about the Ottoman Empire has been manually annotated following these guidelines, creating a gold-standard dataset to test this approach. Furthermore, this study incorporates domain-specific prompt creation to leverage advanced large language models (LLM) to predict the relationships between extracted entities autonomously. By designing customized prompts that align with the linguistic and contextual nuances of historical travelogues, the study enhances the ability of LLMs to accurately identify and classify person-place interactions, action tags, and other relevant relationships. This approach improves the precision and recall of the models and ensures that the predictions are contextually relevant and aligned with the research objectives. Using domain-specific prompts represents a strategic adaptation of LLMs to the unique challenges of historical text analysis, enabling more effective and reliable automation of complex annotation tasks. It is to be mentioned that LLMs, like most neural network models, are black boxes. While several studies discuss the explainability aspect of LLM models, they are not part of this study. Finally, this study represents a foundational effort to integrate the modeling and analysis of place and person entities within narrative texts, providing historians and researchers with a tool to identify and validate historical travel routes accurately. The results demonstrate the robustness of the annotation framework, as evaluated through precision, recall, and F1 scores, with specific models achieving high levels of accuracy.

Contents

1 Introduction **1**
- 1.1 Background and Motivation . . . 2
- 1.2 Research Objectives . . . 5
- 1.3 Thesis Outline . . . 6

2 Related Work **8**

3 Challenges **13**
- 3.1 Challenge 1: Determining Valid Travel Instances . . . 13
- 3.2 Challenge 2: Spatial Relationship Ambiguity . . . 14
- 3.3 Challenge 3: Character Disambiguation . . . 15
- 3.4 Challenge 4: Journey Segmentation and Itinerary Structuring . . . 15
- 3.5 Challenge 5: Implicit and Indirect Travel Indicators . . . 16

4 Foundation **18**
- 4.1 Annotation . . . 18
  - 4.1.1 The MATTER Cycle . . . 18
  - 4.1.2 Scoping the Phenomena . . . 20
  - 4.1.3 Theoretical Models . . . 21
  - 4.1.4 Developing a Preliminary Model . . . 23
- 4.2 Entity Types . . . 23
- 4.3 Named Entity Recognition (NER) Models . . . 24
  - 4.3.1 SpanMarker . . . 24
  - 4.3.2 spaCy . . . 25
- 4.4 Data Extraction Models . . . 25
- 4.5 Large Language Models (LLM) . . . 26
- 4.6 Evaluation Matrics . . . 28

5 Annotation Guidelines **31**
- 5.1 Task Objective, Details and Setup . . . 31

5.2 Annotation Classes 32

6 Data Preparation **60**
6.1 Data and their Digital Format 60
6.2 Data Extraction Procedure 62
6.3 Pre-processing Data 63
6.4 Corpus Preparation 64
6.5 Post-processing Data 67
6.6 Manual Annotation 68

7 Implementation **71**
7.1 Technical Setup and Environment 73
7.1.1 Development Environment 74
7.1.2 Dataset 74
7.2 Integration of Annotation Guidelines 75
7.2.1 Prompt Design 75
7.2.2 Annotation Data Structure 76
7.2.3 Handling ambiguous narratives 78
7.3 OpenAI GPT-Based Annotation Module 79
7.3.1 Methodology for Using OpenAI GPT Models 79
7.3.2 Few-Shot Learning Approach 80
7.3.3 Implementation Details 80
7.3.4 Key Challenges and Solutions 82
7.4 LLaMA-Based Annotation Module 83
7.4.1 LLaMA Model and Integration Process 83
7.4.2 Two-Shot Learning Approach 83
7.4.3 Implementation Details 84
7.4.4 Differences in Approach from OpenAI GPT Model 84
7.4.5 Observed Strengths and Limitations 86
7.5 Mistral-Based Annotation Module 87
7.5.1 Mistral Model and Integration Process 87
7.5.2 Two-Shot Learning Approach 88
7.5.3 Implementation Details 88
7.5.4 Differences in Approach from OpenAI GPT Models 89
7.5.5 Observed Strengths and Limitations 90
7.6 Inference Pipeline and Automation 91
7.6.1 Overall Workflow from Input Text to Annotated Output 91
7.6.2 Error Handling Strategies 93
7.7 Post-possessing Data: 93
7.8 Summary 95

8 Evaluation **97**
8.1 Evaluation Framework . . . 97
8.2 Results . . . 102
8.2.1 Results Discussion . . . 103
8.2.2 Few-Shot Learning Performance . . . 105
8.2.3 Itinerary Comparison . . . 106
8.2.4 ATO Results and Discussion . . . 108
8.3 Error Analysis . . . 109
8.3.1 Limitations in Model Capabilities for Complex Annotation Tasks . . . 109
8.3.2 Inability to Maintain the Specified Output Structure . . . 110
8.3.3 Failure to Detect Historical Place Entities . . . 111
8.3.4 Omissions of Action Tags . . . 111
8.3.5 Introduction of Unauthorized Tags and Invalid Spatial Relation Values . . . 111
8.3.6 Modifications or Omissions of Original Text . . . 112
8.3.7 Errors in Numbering Person Identifiers . . . 112
8.4 Discussion . . . 112

9 Conclusion **115**
9.1 Limitations . . . 116
9.2 Future Work . . . 116

Bibliography **119**

List of Figures

1.1 Agreements between (M)Large Language Models (LLM) and the human annotator groups (i.e., A1 and A2). The compare agreements are based on i) matching either A1 or A2 and ii) inter-annotator agreement between human annotators (A1 vs. A2) and between LLMs and the human majority vote. In the A1 or A2 column, use the same human majority vote to measure the agreements for human annotators. Results are reported separately for English and German. For human annotations, report the total time and cost. Generative Pre-training Transformer (GPT)-4o is used in all steps of the LLM annotation pipeline. [7, p 4] 4

4.1 The MATTER methodology 19
4.2 The MAMA sub-cycle 20
4.3 A diagram illustrating the three steps of instructGPT: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. [32, p 3] 27

7.1 Domain Specialization of LLM 72
7.2 Illustration of standard vs. few-shot inference in OpenAI GPT-based annotation. . 80
7.3 Overall Workflow for Annotating Travelogues Using LLM Models 92

8.1 F1 Scores Comparison Across LLM Models. 103
8.2 F1 Score Comparison: GPT-4o (Few-Shot vs. Two-Shot). 106
8.3 Performance Comparison of GPT-4o vs. LLaMA3:8B 108

List of Tables

5.1 Annotation Classes 32
5.2 Short description of spatial relation's subtask 37
5.3 Examples of ambiguous instances 44
5.4 Short description of travel connection's subtask 54

6.1 Data Source List 61
6.2 Travelogue Corpus 67

7.1 Key challenges and solutions in OpenAI GPT-based annotation. 82
7.2 Comparison of LLaMA (Ollama) vs. OpenAI GPT-Based Annotation 85
7.3 Comparison of Mistral vs. OpenAI GPT-Based Annotation 90

8.1 Error Types and Explanations 99
8.2 Evaluation Schemas and Explanations 100
8.3 Scenarios with Golden Standard, System Predictions, and Evaluation Schema Results 100
8.4 Evaluation Metrics 102
8.5 LLMs F1-Scores Comparison (Two-Shot Learning) 103
8.6 Itinerary Comparison 107
8.7 Comparison of LLM Performance on ATO Dataset (F1-Scores) 108
8.8 Comparison of Original vs. Generated Data 110
8.9 Comparison of LLMs Generated Data 111

List of Abbreviations

ACE	Automatic Content Extraction	23
AI	Artificial Intelligence	2
ATO	Academic Tourism Ontology	74
CIDOC	International Committee for Documentation	
CRM	Conceptual Reference Model	
GTA	Geographical Text Analysis	9
GPT	Generative Pre-training Transformer	x
IE	Information Extraction	9
ISO	International Organization for Standardization	9
LLM	Large Language Models	x
MAMA	Model-Annotate-Model-Annotate	19
MATTER	Model Annotate Train Test Evaluate Revise	19
ML	Machine Learning	2
MUC	Message Understanding Conference	100
NER	Named Entity Recognition	13
NLP	Natural Language Processing	2
OCR	Optical Character Recognition	25
TEI	Text Encoding Initiative	

Chapter 1

Introduction

Historical travelogues, rich with cultural and geographical information, present challenges in ensuring itinerary accuracy, as they frequently mention places not physically visited by the traveler. For example, the following travelogue entry references *"Spanish towns"* and *"St. Philip's castle"* are not part of the actual itineraries since the traveler did not visit these locations. Such mentions often appear in travelogue and unstructured text, as illustrated in example 1.

e.g. 1: *"The Angelica, being freighted for Genoa and Leghorn, we now shaped our course for the former port. We were becalmed, on the 7th of July, near the coast of Spain, off Cape de Gatte. We then had heavy showers and hard gales, by which we were driven out of our way and our masts endangered. Light airs and clear weather followed; the sky was blue and spread with thin, fleecy clouds. We had a view of several Spanish towns and of St. Philip's castle in the Island of Minorca."* [1, p. 5]

Previous methodologies **proposed in the past literature** can identify person and place entities in the text but often **struggle to determine** whether a location was actually visited by the traveler or its relevance to a specific segment of the journey.

This study addresses this limitation by **introducing a classification task to identify travel connections between person and place entities**. This task aims to classify and distinguish between two distinct scenarios:

1. *Instances where an individual's presence at a specific location is directly associated with their journey or travel trajectory* and
2. *Instances where their presence at the location is independent of any journey-related context.*

Finely crafted annotation guidelines are developed to support this task. The guideline contains instructions to systematically label relationships between entities (person and places) while considering their travel connections throughout the travelogue. This guideline is then used as an initial model prompt to perform the desired operation.

1.1 Background and Motivation

Recognizing itineraries within travelogues is critical for historians and researchers, as it provides a clear timeline of events and places. This clarity aids in tracing the sequence of the traveler's journey and contextualizing the historical significance of each visited location. Although advances in Natural Language Processing (NLP) have progressed in relevant areas, including entity tagging, which identifies and categorizes text spans corresponding to entities by type [2]. Comprehensive schemas have been developed to describe relationships between characters and places in narratives [3]. Sandra et al. proposed techniques for interlinking textual data with semantic annotation and ontology design patterns [4]. Nevertheless, a notable limitation persists throughout the literature; while **existing approaches can effectively recognize *person* and *place* entities or establish links between them in unstructured texts, they often strive to ascertain if a particular place was visited by the traveler or its significance to a particular segment of the journey**. This highlights the challenge of interpreting contextual nuances and inferring implicit information, such as the relevance of a location within the broader narrative of the trip.

As an alternative, well-defined travel-specific annotation guidelines can provide a robust framework for analyzing historical travelogues. **By integrating multiple existing annotation classes with additional travel-specific categories**, such as *journey segments*, *connections between characters*, and *travel-related connections*, the framework can enhance consistency and reliability in the annotation process. This approach ensures that the unique characteristics of historical travel narratives such as **spatial movements, temporal progressions, and contextual interactions** are accurately captured and represented. The inclusion of travel-related classes not only refines the granularity of annotations but also reduces ambiguity, improving data quality and facilitating more precise analysis [More details on this can be found in Chapter 8]. Such a comprehensive framework supports the extraction of meaningful insights from historical travelogues while maintaining operational efficiency and minimizing bias. **Ultimately, such guidelines are critical for ensuring accurate and reproducible analysis of historical narratives, facilitating more reliable insights and conclusions.**

Furthermore, annotating data based on annotation guidelines involves labeling or tagging data to enhance its interpretability and utility for Machine Learning (ML) and Artificial Intelligence (AI) models. This process entails connecting descriptive labels to raw data to provide context, which aids ML and AI models in their learning and decision-making processes. **Data annotations enable machines to identify and differentiate entities within the data by supervised training.** This, in turn, allows them to map different relationships within given texts, allowing them to understand different contexts and relationships. For instance, in text annotation, texts are supplemented with additional information, labels, and definitions to clarify their meaning and structure. In image annotation, objects and entities within an image are highlighted and outlined to make the content interpretable for machines. The quality of annotated data is critical, as it directly

influences the accuracy, reliability, and overall performance of ML and AI models [5, p. 1]. High-quality annotations ensure that models can generalize effectively and produce meaningful insights, underscoring the importance of meticulous and consistent annotation practices.

Additionally, these guidelines help to minimize the cost and time involved with manual annotation through multiple mechanisms. Data annotation can either be manually performed by human annotators or automatically performed using sophisticated ML and AI techniques with less to no human intervention or supervision. Nonetheless, manual annotation can be laborious and time-intensive, posing a significant challenge to ML and AI implementation for organizations. Despite these challenges, the importance of data annotation in the fields of AI and ML remains undeniable. It serves as the foundation of these technologies, empowering models to comprehend and interpret data, identify patterns, and make precise predictions. **This hybrid approach of manual expertise combined with automated tools** can save considerable time and money while providing the same data quality. For instance, Large Language Models (LLM) are advanced AI systems designed to understand, generate, and manipulate human language. **By integrating human expertise with the capabilities of Large Language Models (LLM)s**, this hybrid approach leverages the strengths of both manual and automated methods. Human annotators provide nuanced understanding and domain-specific knowledge, while Large Language Models (LLM)s offer scalability and efficiency in processing large volumes of data. Together, they ensure high-quality annotations, reduce the time and cost associated with purely manual processes, and enhance the overall accuracy and reliability of the data [Figure 1.1]. This synergy between human intelligence and machine automation represents a powerful strategy for optimizing data annotation workflows, particularly in complex or domain-specific tasks.

The above discussion leads us to the research questions, which underscore the necessity of a meticulous and systematic approach guided by two primary inquiries:

1. *How to identify valid itineraries within historical travelogues?* and
2. *How to leverage LLMs for complex, multi-step reasoning task?*

Identifying valid itineraries of travelers' journeys in historical travelogues presents significant challenges due to these narratives' inherent ambiguities and subjective nature. Such travelogues offer valuable insights into travelers' experiences, observations, and routes [6, p. 7-10]. Scholars and researchers often seek to delineate travel paths within these texts to reconstruct the chronology of events and assess the significance of each visited location. Tracing these itineraries can give a deeper understanding of each destination's historical context and relevance. **This research underscores the importance of addressing the central question: *How to identify valid itineraries within historical travelogues?*** It emphasizes the need for a meticulous and systematic approach to navigate the complexities of these texts, ensuring accurate and meaningful reconstruction of historical journeys.

Additionally, the study proposes **leveraging modern state-of-the-art LLM to implement annota-**

tion guidelines that facilitate the investigation and validation of itineraries in historical travelogues. The advent of LLM has transformed the landscape of manual annotation, altering it to be fast and inexpensive. Automated pre-annotation early-stage sorting can be performed on large datasets by these models, estimated to be 100 to 1,000 times more cost-effective than human annotators [7]. The use of LLMs to detect uncertain or complex cases would substantially reduce the amount of data where human review is required; **Hosseini et al estimated that when using LLMs, only 35% of the data needed human annotation, leading to a 62% cost reduction and 63% decrease in time spent on labor as shown in figure 1.1 [7, p 4].** LLMs decrease the dependency on human annotators so that large datasets can be annotated faster and more cost-efficient than workflows that rely exclusively on humans; thus, they are particularly suited for significant data initiatives.

Annotator	Agreement with A1 OR A2 (%)		Inter Annotator Agreement (%)		Total time		Cost	
	EN	DE	EN	DE	EN	DE	EN	DE
A1	76.6	75.9	60.2	60.5	3 weeks		€15,000	
A2	75.8	76.1	60.2	60.5	3 weeks		€15,000	
LLM-text	75.4	75.6	62.8	61.4	8 min	7 min	€70	€70
MLLM-text	76.6	77.0	64.3	63.1	12 min	13 min	€92	€93
MLLM-multi	**78.1**	**78.4**	**65.6**	**64.7**	10 min	10 min	€72	€72
MLLM-multi (gpt-4-turbo)	75.4	75.6	62.9	62.0	30 min	23 min	€155	€156
Without query-specific annotation guidance								
LLM-text	71.2 (**-4.2**)	67.0 (**-8.6**)	57.6 (**-5.2**)	52.2 (**-9.2**)	7 min	7 min	€45	€45
MLLM-text	72.5 (**-4.1**)	68.0 (**-9.0**)	58.7 (**-5.6**)	53.4 (**-9.7**)	11 min	11 min	€66	€67
MLLM-multi	74.1 (**-4.0**)	71.9 (**-6.5**)	61.6 (**-4.0**)	59.9 (**-4.8**)	4 min	4 min	€25	€24

Figure 1.1: Agreements between (M)LLM and the human annotator groups (i.e., A1 and A2). The compare agreements are based on i) matching either A1 or A2 and ii) inter-annotator agreement between human annotators (A1 vs. A2) and between LLMs and the human majority vote. In the A1 or A2 column, use the same human majority vote to measure the agreements for human annotators. Results are reported separately for English and German. For human annotations, report the total time and cost. GPT-4o is used in all steps of the LLM annotation pipeline. [7, p 4]

Moreover, LLMs produce more consistent annotations over large datasets, which reduces the need for multiple human annotators, thus decreasing conflicts in labeling [8]. **Though they may not reach expert-level quality in the most specialized domains, LLMs greatly expedite the annotation process, making it much faster and more cost-effective at scale, thereby reshaping the future of data annotation in NLP and related fields.** However, this study acknowledges that LLM models typically function as "black boxes," generating responses from input data without revealing the rationale behind these responses. These models' lack of transparency and interpretability presents a notable limitation, potentially leading to various challenges and issues, which is another area of research [8]. This study does not cover that area as **the main purpose of this research is to provide annotation guidelines that help researchers retrieve valid itineraries within historical travelogues.**

1.2 Research Objectives

The primary aim of this research is to integrate place and person entity modeling and analysis by examining their travel connections within narrative texts, supported by the introduction of travel-specific annotation guidelines. By developing a systematic framework for identifying and annotating entities such as travelers, locations, and their interactions, this study seeks to enhance the accuracy and consistency of extracting meaningful insights from historical travelogues. These guidelines will address the complexities of narrative texts, such as ambiguities in entity relationships and the subjective nature of historical accounts, enabling a more structured and reliable approach to analyzing travel-related data. Ultimately, these guidelines provide historians and researchers with a robust tool for accurately identifying and validating historical travel routes.

Furthermore, to evaluate the proposed annotation guidelines, a corpus of 120 pages from a collection of 19th-century English-language travelogues on the Ottoman Empire has been manually annotated to establish a gold-standard reference [9]. Additionally, this study incorporates domain-specific prompt creation to leverage advanced LLM models to predict the relationships between extracted entities autonomously. By designing customized prompts that align with the linguistic and contextual nuances of historical travelogues, the study enhances the ability of LLMs to accurately identify and classify person-place interactions, action tags, and other relevant relationships. This research aims to advance the field of historical text analysis by providing a fundamental methodology for reconstructing and interpreting travel itineraries and their contextual significance.

The corpus and custom models based on advanced LLM developed through this study are intended to function as versatile research tools in areas such as historical travel studies and digital humanities. By systematically annotating historical travelogues, this work seeks to enhance both the usability and accessibility of these documents, paving the way for more in-depth scholarly investigations. Furthermore, this research also highlights the challenges and issues posed by LLMs.

The primary aims of the research are as follows:

1. Assemble and prepare a corpus of 120 pages from several 19th-century English historical travelogues that align with the study period and itineraries of interest.
2. Develop and validate comprehensive annotation guidelines to identify valid travel itineraries.
3. Annotate the travelogue corpus using state-of-the-art Large Language Models (LLMs) in compliance with the established annotation guidelines.
4. Assess the effectiveness of the annotations and refine both the model and annotation procedure through performance metrics such as precision, recall, and F1 score.

1.3 Thesis Outline

This report is organized into nine chapters, each addressing a critical aspect of the research. The introductory chapter provides an overview of the topic, including its background, motivation, and objectives. **Relevant literature** is reviewed in *chapter 2*, emphasizing recent studies and their contributions to the field. The **key challenges of existing approaches** are presented in chapter 3. **Foundational terminology and concepts** related to annotation, data extraction, and implementation are explored in *chapter 4*. The proposed **annotation guidelines** are detailed in *chapter 5*, supplemented with illustrative examples of both correct and incorrect implementations. **Data format, extraction processes, and procedures** for corpus preparation and manual annotation are outlined in *chapter 6*. A comprehensive description of the **implementation process**, including basic statistics of the dataset and an individual analysis of each model, is provided in *chapter 7*. The **evaluation methodology** is discussed in *chapter 8*, presenting the metrics used, the results obtained, and an in-depth analysis of the findings. Finally, *chapter 9* **summarizes the overall contributions** of the research, identifies limitations, and proposes potential directions for future work.

Chapter 2

Related Work

This chapter provides an overview of relevant research that forms the foundation for this study's goals. Previous studies have explored various linguistic and computational methods, from identifying general expressions to annotating causal, temporal, and spatial connections. While numerous researchers have suggested techniques for categorizing and linking entities, few have comprehensively tackled distinguishing between authentic travel itineraries and referenced locations in historical travel accounts.

Ananieva et al. established the **"Digital Editions of Historical Travelogues" (DEHisRe)** project, which employs a six-step "Life Cycle of Digital Editing" to convert handwritten manuscripts into digital scholarly editions [10]. By utilizing standards like TEI and CIDOC CRM, in conjunction with GPT-4 for improved entity recognition, the project facilitates the visualization and monitoring of travel narratives. Nonetheless, the challenge of **verifying a traveler's actual itinerary** remains unresolved, as the current methodology primarily focuses on broad entity identification and visualization without explicitly establishing travel relationships between them.

Sandra et al. outline how **semantic annotation and ontology design patterns** can enhance the analysis of historical travelogues by encoding key travel features such as time, space, and events [4]. Drawing on standards like CIDOC CRM and TEI, the authors aim to capture travel dynamics with greater expressivity, facilitating advanced queries (e.g., via SPARQL) and visualizations (e.g., GIS-based). By integrating widely accepted frameworks with project-specific elements, DEHisRe seeks to bridge curation-driven and research-driven perspectives, offering a flexible approach to modeling and retrieving historical journey data. Although this method significantly enhances the organization and retrieval of place and character data, it **does not definitively differentiate between destinations the traveler physically visits and those merely mentioned incidentally**.

A framework for classifying **spatial relationships between characters and places in narrative texts** was proposed by Soni et al. [3]. This framework distinguishes whether a character is situated in, near, moving toward, or moving away from a specific place, thereby shedding light on narrative

progression and character development. While these spatial relationships offer insights into narrative structure, further analysis is necessary to ascertain if these locations serve as valid waypoints along a documented journey. The proposed annotation guidelines have **incorporated this concept to establish the spatial association between a character and a place reference** [Chapter 5.2]. The categorization of spatial relationships, in conjunction with the principles of SpatialML, has been applied within the annotation guidelines as a **Spatial Relation class** to systematically determine the spatial relationship between a character and a place. This classification later aids in constructing a valid itinerary [Section 5.2] [11].

A multilayer semantic annotation protocol based on International Organization for Standardization (ISO) standards to extract information from European Portuguese newswire articles was introduced by Silvano et al. [12]. This method incorporates layers for temporal, referential, and semantic roles. While demonstrating the adaptability of a multilayer system to different text types, this protocol **does not explicitly address whether characters physically visit the mentioned places**, a critical consideration for historical travel accounts.

For analyzing historical documents like the Relaciones Geográficas de la Nueva España corpus, **geographically focused techniques in NLP and ML** were developed by Bara et al. [13]. Their Geographical Text Analysis (GTA) methodology automatically recognizes place names and associates them with real-world coordinates, easing the identification of geospatial patterns in historical data. Despite this advancement, **the approach does not confirm whether the identified locations are part of a traveler's route**.

Ardanuy presented strategies for **disambiguating two types of named entities, toponyms and person names**, in historical digitized texts, emphasizing entity-centric text mining [14]. The GeoSem method proposed for toponym resolution integrates geographic and semantic features from knowledge bases such as Wikipedia and GeoNames, improving upon state-of-the-art toponym resolution benchmarks in English, German, and Dutch corpora. A separate method, SNcomp, approaches person name disambiguation as a dynamic clustering problem and returns automatically generated social networks for disambiguated entities. While **these approaches greatly aid in clarifying references to places and individuals, they do not address whether identified places are integral to a traveler's actual itinerary**.

Luan et al. established a general **Information Extraction (IE) framework that tackles key IE tasks** in a unified manner [15]. Their approach, called "Dynamic Span Graphs," jointly models entity recognition, relation extraction, and coreference resolution by enumerating all potential text spans and iteratively refining these representations using global context. Furthermore, it can handle nested or overlapping entities—an essential consideration for domain-specific annotation. This method achieves state-of-the-art results across multiple domains without heavy reliance on external syntactic tools. Despite their person-place and person-person relationship extraction, the approach is not sufficient for itinerary extraction from historical travelogues, as it **does not distinguish**

different types of relationships between person entities—and such travelogues typically contain multiple person entities who may or may not have a travel connection, which leads to false positives in journey segmentation. Although these methods effectively clarify references to locations and individuals, they still **do not ascertain whether the identified places are genuinely part of the traveler's itinerary**.

The **CaTeRS (Causal and Temporal Relation Scheme) framework**, which annotates various causal and temporal connections between events in narratives, was introduced by Nasrin et al. [16]. The system enhances the understanding of event sequences by classifying these links into nine causal and four temporal categories. This type of annotation aids in grasping how stories evolve over time; however, additional criteria are necessary to confirm if a location is essential to a traveler's route rather than just a passing reference. **Temporal relationships have been incorporated into the proposed annotation guidelines as the Temporal Span class to distinguish between a character's short-term and long-term association with a location** [Section 5.2].

Reiter and Frank developed a **supervised machine learning approach to identify generic noun phrases, which convey general truths or rule-based knowledge instead of precise instances** [17]. Distinguishing these phrases from specific entities refines the process of extracting information. In the context of historical travel accounts, filtering out generic references ensures a more accurate identification of truly pertinent entities. **This concept has been incorporated into the proposed annotation guidelines as the Valid Relation class to filter out generic person and place entities, ensuring a more precise association between characters and locations** [Section 5.2].

Mani et al. proposed **SpatialML**, which aims to standardize the tagging of place mentions using labels like *PLACE*, *LINK*, and *PATH* [11]. These tags promote compatibility with geographical standards and databases. Although the framework can link mentions to specific geographic coordinates, it does not specify if the traveler actually visited these coordinates as part of a coherent journey. **The concept of SpatialML has been integrated into the proposed annotation guidelines as a Spatial Relation class to determine the spatial relationship between a character and a place, thereby facilitating the identification of a valid itinerary** [Section 5.2].

In conclusion, previous research has developed robust tools for detecting entities, annotating spatial and temporal elements, and semantically structuring narrative texts, all of which have significantly contributed to historical document analysis. These advancements have improved entity recognition and spatial mapping techniques, facilitating structured representations of historical texts. However, despite these progressions, most existing approaches struggle to differentiate between locations that travelers genuinely visited and those that are merely referenced within the narrative. This limitation poses a significant challenge in reconstructing authentic itineraries, as conventional methods often misattribute spatial relationships without considering the actual movement of the traveler.

To address this gap, the current study introduces a **travel-specific classification task** that focuses

on establishing explicit travel connections between characters and locations. By incorporating linguistic, spatial, and temporal markers, this approach systematically distinguishes between places that are part of a traveler's journey and those that serve only as contextual references. This refined methodology enhances the identification of authentic itineraries within historical travel narratives, enabling a more structured and accurate reconstruction of historical movements.

Furthermore, by refining the distinction between visited and referenced locations, this research contributes to improving the **precision and reliability of historical text analysis**. The findings not only aid in reconstructing valid itineraries but also provide a deeper understanding of travel patterns, character interactions, and the broader socio-historical context embedded within historical travelogues.

Chapter 3

Challenges

This chapter presents and discusses the key challenges posed by previous methodologies.

The study identifies a fundamental challenge in **valid itinerary retrieval**, necessitating additional labeling mechanisms to accurately recognize valid itineraries and establish travel connections between **Person (PER)** and **Place (PLACE)** entities within historical travelogues. While current **Named Entity Recognition (NER)** methodologies effectively identify person and place entities within textual data, they cannot ascertain whether the traveler visited a referenced location or if it is relevant to a specific segment of the journey.

A critical limitation of existing approaches is their inability to differentiate between the **primary traveler's movements** and **background references** to other locations. Travel narratives often contain **past references**, **hypothetical journeys**, and **secondary character movements**, making distinguishing between actual travel events and contextual descriptions challenging. Additionally, the system must account for **multiple travelers** within the same passage, requiring a robust mechanism to classify and associate journey-specific information with the correct entity. At a more granular level, the study identifies the following key challenges:

3.1 Challenge 1: Determining Valid Travel Instances

One of the fundamental challenges in travelogue annotation is distinguishing between actual travel events and non-travel references. Many texts contain place mentions that do not indicate movement or visitation. These include:

- **Hypothetical Journeys:** Future travel intentions, planned trips, or conditional statements.
- **Past Travel Mentions:** Locations associated with previous journeys, not part of the current itinerary.
- **Travel Mentions of Other Characters:** References to journeys undertaken by individuals

other than the main traveler, which may incorrectly be attributed to the protagonist (More details on this can be found in Section 3.3).

For example, a sentence such as *"John Stewart took leave of England for America [18]"* does not imply that *John reached America* but indicates a future reference to America. Similarly, *"The plague, as we were informed at Leghorn, having appeared at this place in the spring, our captain was unwilling to arrive there before it should have ceased [1]"* refers to an *past journey* rather than a current travel event.

Furthermore, in the sentence *"Our party was increased by the arrival of the Honble. Captain Irby and Captain Mangles, both of the Royal Navy, who had been absent for a short time to visit Bethlehem. [19]"*, the mentioned *location—Bethlehem—refer to Captain Irby and Captain Mangles's journey, not the main travelers (whose journey is of primary interest in the story)*. If such instances are incorrectly included in itinerary extraction, they could lead to misleading historical travel representations. This highlights the importance of distinguishing between related travel events and unrelated references to ensure the accuracy and reliability of historical journey analysis.

3.2 Challenge 2: Spatial Relationship Ambiguity

In many cases, a location reference does not explicitly indicate movement, making it difficult to determine whether a traveler actually visited a place. Additionally, spatial relationships may link locations to individuals who are not part of the journey, further complicating itinerary reconstruction. This ambiguity arises in four main scenarios:

- **Temporary stays**, where a traveler resides in a place for an extended period but does not engage in active travel.
- **Past references** to locations visited in previous journeys that are unrelated to the current travel context.
- **Future or intended travel**, which describes plans or aspirations without confirming actual movement.
- **Link to unrelated individuals**, where a location is mentioned in association with a person who is not the primary traveler, leading to potential misattributions.

For example, the sentence *"He stayed in Cairo for several months"* suggests that the traveler spent time in *Cairo* but does not indicate movement from another location. Similarly, *"I again changed my plan and decided on continuing my journey by land to Roustchouk [20]"* conveys future intent rather than completed movement.

In the case of unrelated individuals, a sentence like *"I was requested to visit an Armenian merchant, who had just arrived from Damascus [21]"* references to *Damascus*, but the primary traveler may have no direct connection to *Damascus*. If such mentions are not carefully analyzed, they may introduce *false travel links* in itinerary reconstruction. Traditional entity recognition methods struggle to

distinguish travel-related actions from non-travel associations, often leading to false positives in journey segmentation.

3.3 Challenge 3: Character Disambiguation

Another critical challenge is ensuring that the correct character is associated with a travel event. Travelogues often mention multiple individuals, making it essential to differentiate between:

- The **main traveler**, whose journey is the primary focus of the narrative and whose movements should be systematically tracked to reconstruct the itinerary accurately.
- **Companions**, who may accompany the main traveler on certain segments of the journey but are not necessarily the central figure of the travelogue.
- **Unrelated individuals**, whose movements and locations are mentioned in the text but are independent of the main traveler's journey and should not be linked to the primary itinerary.

For example, in the sentence *"Alex and Maya departed from Vienna"*, both characters are directly involved in the movement, implying a shared travel event. However, in *"He met Professor Eric, who had traveled from Istanbul"*, the past journey of *"Professor Eric"* is unrelated to the main traveler's current movement. If entity recognition models fail to differentiate between relevant and irrelevant character movements, the resulting annotations may mistakenly associate *unrelated locations* with the primary traveler's itinerary.

Such misattributions can distort historical itinerary reconstruction by introducing incorrect travel paths or implying interactions that did not occur. In cases where multiple individuals are mentioned in a travelogue, ambiguity arises if the model does not correctly identify the *primary character of interest* and their direct travel connections. Additionally, the presence of companions further complicates annotation, as *group travel events* must be distinguished from individual movements. These challenges highlight the necessity of *context-aware entity recognition models* to ensure accurate representation of travel sequences.

3.4 Challenge 4: Journey Segmentation and Itinerary Structuring

Accurately segmenting a traveler's journey involves identifying:

- The **starting point** of the journey.
- **Intermediate stops** indicate places passed through or temporarily stayed in.
- The **final destination**, marking the endpoint of a travel sequence.

Consider a narrative describing a journey from Venice to Prague via Vienna: *"He departed from Venice, passed through Vienna, and arrived in Prague."* A well-structured annotation system must

ensure that these locations are ordered sequentially in the itinerary: *Venice* $\rightarrow$ *Vienna* $\rightarrow$ *Prague*.

However, challenges arise when travel descriptions are *non-linear or incomplete*. Some narratives may introduce detours, prolonged stays, or indirect references, such as: *"After spending weeks in Athens, he later found himself wandering the streets of Istanbul."* Although travel is implied, the direct transition between these locations is missing, making segmentation difficult.

3.5 Challenge 5: Implicit and Indirect Travel Indicators

Not all travel movements are explicitly stated. In many cases, a location is mentioned in *implicit ways*, requiring contextual inference to determine whether the traveler actually visited that place.

- **Indirect Travel Mentions:** *"After months in Athens, he found himself in Istanbul."* There is no explicit mention of traveling between Athens and Istanbul, yet the phrase *"found himself in Istanbul"* suggests movement.
- **Non-Travel References:** *"He read extensively about Alexandria in preparation for his trip."* Despite the mention of Alexandria, there is no evidence that the traveler ever went there.

Such cases pose a significant challenge for standard entity recognition systems, requiring *context-aware inference* rather than simple entity tagging.

These challenges underscore the complexity of annotating travelogues, where historical, linguistic, and contextual factors must be carefully considered. Conventional NER methods are inadequate for accurately identifying travel-related entities and structuring itineraries, as they struggled to capture the nuanced relationships between characters, locations, and movement. Overcoming these limitations requires:

- *Travel-based spatial reasoning techniques* to distinguish between valid and non-valid place references.
- *Character disambiguation models* to accurately associate travel actions with the primary traveler rather than unrelated individuals.
- *Context-aware inference mechanisms* to detect implied movements and infer travel sequences even when explicit travel actions are absent.

Effectively addressing these challenges necessitates the development of a **structured annotation framework** that extends beyond conventional entity recognition. This framework should integrate **spatial reasoning**, **temporal classification**, and **context-aware journey modeling** to improve itinerary retrieval and ensure accurate representation of travel sequences in historical narratives.

Chapter 4

Foundation

This chapter introduces and explains the fundamental principles and methods of this research. Concepts such as the MATTER cycle, entity types, and data extraction models have been pivotal in shaping the approach and deriving the proposed solution. Subsequent chapters will discuss these foundational elements in detail, comprehensively understanding their contributions to the research framework and methodology.

4.1 Annotation

Linguistic annotation entails associating descriptive or interpretive labels with language data in various formats such as plain text, audio files, video captures, and physiological recordings [22, p 22]. These labels encompass transcriptions, part-of-speech tags, syntactic structures, and semantic annotations. Some resources offer single annotations, while others present multiple layers of annotations. The annotation scheme is central to linguistic annotation projects, delineating the labels and features for specific linguistic units. Established annotation schemes can be leveraged for known phenomena, or projects may opt to devise unique classes of annotations when exploring new linguistic phenomena. Developing novel schemes typically involves iterative annotation, evaluation, and refinement processes. The annotation methodology and concept used to develop the presented annotation guidelines [Chapter 5] are discussed as follows:

4.1.1 The MATTER Cycle

The main goal of an annotation project is to ensure that the selected features used to represent a specific phenomenon are detailed enough to effectively train ML algorithms for automatic classification in a particular task [22, p 22-24]. Linguistic descriptions are based on comprehensive theoretical models of a phenomenon. However, practical tasks involve modeling a single linguistic phenomenon and understanding the interaction between multiple independent phenomena from a theoretical perspective. As a result, a ”theoretically informed” annotation model is not purely

linguistic. These descriptions form the basis for the annotation values in the specification language, which are then utilized as features in a development cycle to train and test text identification or labeling algorithms. [Fig. 4.1].

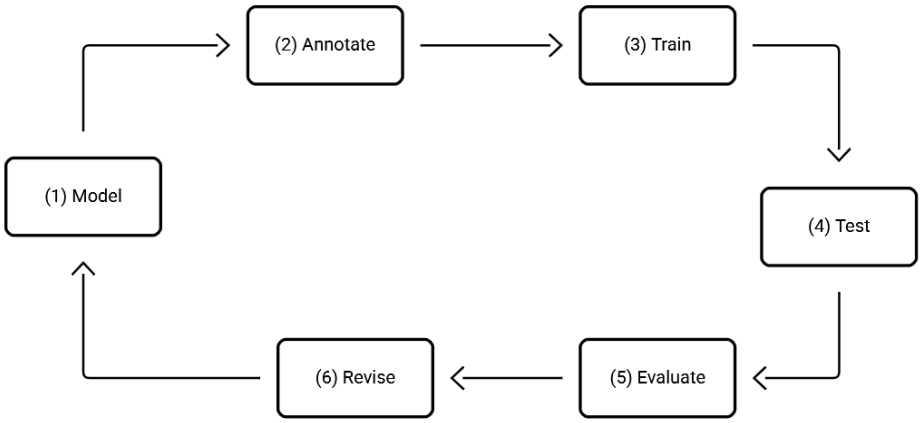


Figure 4.1: The MATTER methodology

Subsequently, after analyzing and evaluating system performance, the phenomenon model may undergo revisions for further training and testing. This development cycle, known as the Model Annotate Train Test Evaluate Revise (MATTER) methodology, includes the following steps [22, p 23]:

a. **Model:** Structural descriptions are informed by theory and derived from empirical data observations.

b. **Annotate:** The annotation scheme assumes a feature set that encodes the input data's specific structural descriptions and properties.

c. **Train:** The algorithm is trained on a corpus that has been annotated with the target feature set.

d. **Test:** The algorithm is assessed using held-out data.

e. **Evaluate:** Utilize standardized methods for evaluating the results.

f. **Revise:** Revisit the model, annotation specifications, or algorithm to enhance the robustness and reliability of the annotation. Within the MATTER cycle, the model revision process is known as the Model-Annotate-Model-Annotate (MAMA) cycle, or the phase referred to as "babbling," as shown in Figure 4.2.

This research uses the MATTER cycle to develop the annotation guidelines. Following the MATTER cycle, this research continuously improves the annotation model and its application, ensuring

high-quality data for analysis and further system development.

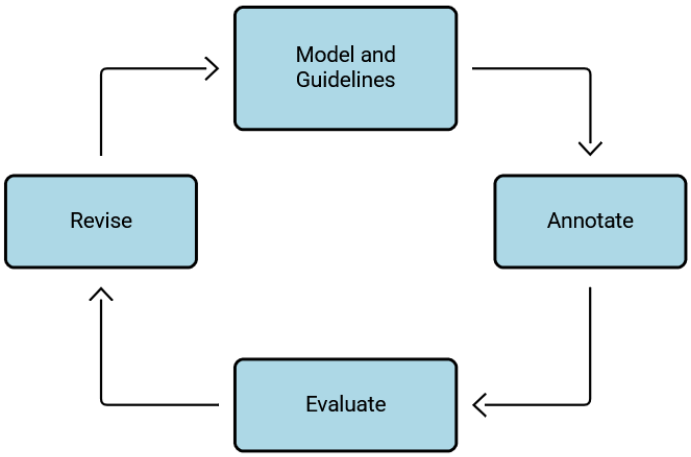


Figure 4.2: The MAMA sub-cycle

4.1.2 Scoping the Phenomena

As natural language understanding systems progress beyond essential keyword recognition and named entity extraction, interpreting temporal and event expressions becomes essential [22, p 27-28]. It involves identifying events and their participants, anchoring them temporally, and structuring them into timelines for reasoning. However, a significant portion of the temporal information in texts is implicit and imprecise. A critical aspect of automatic information extraction is recognizing and explicitly indicating when events take place. While advancements have been achieved using statistical methods, they rely on annotated data, which relies on annotation schemes and models. In the past, there was no comprehensive framework for annotating time and events, partly due to the challenge of translating linguistic theories into practical models. Initial endeavors focused on annotating temporal expressions (e.g., dates, durations) and timestamping domain-specific events. However, these efforts were narrow and did not encompass broader linguistic aspects such as event semantics, tense, or temporal relationships. The establishment of TimeML aimed to fill these voids by formulating an exhaustive specification for temporal information that encompasses event identification, sequencing, and reasoning about event persistence and outcomes [22, p 27-28]. Constructing a robust annotation framework required incorporating insights from event semantics, temporal logic, and linguistics. This initiative applies this framework to develop the annotation categories within the guidelines, ensuring the creation of well-structured and distinctly defined annotation classes that streamline the labeling process. By adopting the methodology, this research sets up a framework for consistency and accuracy in annotation work, generating more dependable data for further analysis and supporting this research's overarching

goals.

4.1.3 Theoretical Models

Events are positioned in time, connected to each other, and have an internal temporal structure, which can involve having a duration or being momentary and may or may not have a clear endpoint [22, p 29]. Furthermore, language users can perceive events as ongoing or completed, as well as actual or hypothetical. The diverse characteristics of events and how they are perceived are frequently articulated through language. These notions are elaborated on in the subsequent section.

a. **Tense and Aspect**

The primary function of verb tense is commonly to indicate the temporal location of events, either in the past or future. For example, English uses the past tense, such as ”Mia ran,” for events that occurred in the past. In contrast, future events are often indicated using modal auxiliaries like ”will,” as in ”Mia will run.” Some languages express tense differently, for instance, by incorporating aspectual markers into noun phrases or, as seen in Mandarin, lacking explicit tense markers altogether [22, p 29]. Time adverbials like ”tomorrow” and ”today” are also used in many languages to convey tense without changing the grammar. English morphologically distinguishes between two verb tenses (present and past) but conceptually includes three, adding the future tense. While tenses are generally related to the speaking time, they can indicate future events or even express epistemic certainty, as in the phrase, ”She will be at work by now.” The complexity of tense and its varying expressions across languages poses a significant challenge for tasks like 'temporal annotation.'

b. **Event Semantics**

An annotation model improves temporal understanding by constructing a detailed representation of events articulated in language. Events are viewed as intricate constructs with distinct components, forming an event ontology that aids in making semantic differentiations in a text [22, p 30-33]. Linguistics offers three main approaches to modeling events: Aktionsarten categorizes predicates based on event types or aspectual classes, Events, as Arguments, treats events as variables for logical deduction, and Typed Event Structure, which combines both perspectives. Vendler's classification of verbs by temporal characteristics distinguishes between processes, states, accomplishments, and achievements, which vary in duration and culmination. For inferential reasoning, Davidson's event-as-argument approach examines the relationship between predicates and events [22, p 30-33]. Recent studies on event decomposition and scalar change further refine our comprehension by exploring how events progress over time and how participants' roles influence their outcomes. These insights guide the accurate analysis and reasoning of annotated events.

c. **Time Expressions**

Languages express time in diverse ways beyond tense and aspect. Time can function as both a subject or a modifier, for example, ”Friday works better” or ”The plane arrived late” [22, p 33-36]. Some time expressions are explicit, such as dates, while others are relative or vague, like ”yesterday” or ”recently.” Duration, such as ”three weeks,” and anchored durations, like ”two weeks from today,” help measure or connect events in time. In contrast, frequencies like ”every Saturday” signal recurring events. Normalizing these expressions is crucial to constructing a coherent timeline since they vary in precision. Challenges emerge from dealing with temporal relations, indexicals, and vague expressions, necessitating meticulous mapping to establish a useful temporal annotation model [22, p 33-36].

d. **Temporal Relations**

After analyzing the characteristics of events, tense, and temporal expressions, the attention turns towards organizing events chronologically and determining their connections. Understanding temporal relationships is essential for grasping causality, making plans, and performing various cognitive functions. Temporal information in natural language is conveyed through tense, aspect, and adverbials, which a model must consider to make inferences about event timing [22, p 36-40]. For instance, in a story, a system for temporal reasoning should be able to establish the order of events (e.g., running preceding falling). Establishing temporal connections between events, time instances, or both is crucial for constructing a timeline. These connections encompass concepts like ”before,” ”after,” ”overlap,” and ”during.” Utilizing temporal logic and interval relations aids in representing and reasoning about these temporal connections. Diverse methods, such as Allen’s interval algebra, offer a structure for encoding time intervals and their relationships (e.g., ”start,” ”finish”). This structure plays a fundamental role in translating temporal relations into a model for linguistic annotation, facilitating reasoning across artificial intelligence, computational linguistics, and planning tasks [22, p 36-40].

e. **Subordinating Relations**

This section discusses how events are positioned relative to aspect, intention, belief, desires, plans, factuality, and other modalities [22, p 41-42]. Initially, the focus was on representing fundamental temporal meanings. However, the current emphasis is on situating events. Phrases like ”Mia might have bought wine” or ”Tom plans to visit Germany” indicate uncertainty regarding whether events have taken place. Conversely, some statements leave no doubt about an event: ”Mia forgot she locked her car.” Conversely, some statements confirm that something did not happen, such as ”Mia forgot to lock her car.” Moreover, veridicity (the truth value of event sentences) plays a role in determining meaning, as seen in sentences like ”The agency reported....” or ”Officials claim....” which may be true or false

depending on the source. These subordinating relationships are not solely temporal and should be captured in annotation models to correctly interpret an event's factual nature and context [22, p 41-42].

Various factors are considered during the creation of the annotation guidelines. In addition, aspects like event hierarchy, modality, and truthfulness ensure that the guidelines capture the nuanced ways in which events are depicted and characterized depending on the situation. This method enables accurate modeling of temporal details and clarifies the internal relationships of events. Consequently, this enhances the accuracy of reasoning and analysis reliant on event annotation.

4.1.4 Developing a Preliminary Model

This section focuses on establishing an initial model and specification for the temporal phenomena identified earlier. This research's annotation guidelines follow the MATTER cycle methodology, as described in Section 4.1.1, with a specific focus on the MAMA sub-cycle. Within this sub-cycle, proposed initial models are validated through small-scale annotation experiments. The iterative modeling and annotation process helps enhance the specification language for the current task. Various theoretical models related to temporal expressions, tense, aspect, event semantics, and temporal relations were examined. Furthermore, past computational research was reviewed to evaluate the implementation status of the overall objectives [Chapter 2].

4.2 Entity Types

In this research, the Automatic Content Extraction (ACE) annotation guidelines are followed for all entity classes, limiting annotations to pronouns, proper nouns, common nouns, and non-generic noun phrases (i.e., excluding generic noun phrases or WH-question words) [23]. The entity annotation, which is based on the ACE guidelines, focuses on a specific set of categories such as people (PER), natural locations (LOC), built facilities (FAC), and geopolitical entities (GPE) [24, p. 2139]. The entity classes are as follows:

PER. A person refers to an individual indicated by a proper name (e.g., John Stewart) or a common noun (e.g., the man), as well as a group of people, such as his daughters or the Buddhas.

FAC. ACE guidelines define a facility as a "functional, primarily man-made structure" intended for human occupancy (such as buildings and museums), storage purposes (like barns and parking garages), transportation infrastructure (including streets and highways), and maintained outdoor areas (such as gardens).

GPE. Geo-political entities represent unified entities comprising a population, governing body, geographical positioning, and defined political borders. In a literary context, these encompass not just cities with identifiable real-world locations (such as London or New York) or nations (like

England or the United States) but also fictional entities, both named and generic (the town, the village).

LOC. Locations refer to entities characterized by physical presence but without political associations. Within the dataset, these encompass named regions lacking political structures (e.g., New England, the South) and celestial bodies (e.g., Mars). Nevertheless, the prevailing category consists of geological regions delineating natural environments, including the sea, rivers, countries, valleys, woods, and forests.

Generic Noun Phrase. A generic noun phrase refers to a kind or class of individuals rather than a specific individual or group. For example, the noun phrase "The lion" is interpreted as representing the general category of lions rather than a particular lion [17, p 40]. Generic noun phrases can be used with predicates related to types, as shown in (a), or with verbs that describe specific actions, as demonstrated in (b). Unlike in (a), the action described by the verb phrase in (b) can apply to individual lions.

a. **The lion** was the most widespread mammal [17, p 40].

b. **Lions** eat up to 30 kg in one sitting [17, p 40].

4.3 Named Entity Recognition (NER) Models

NER plays a crucial role in various NLP applications, such as machine translation, sentiment analysis, and information extraction. Like many other NLP tasks, it is gaining importance due to the widespread use of the Internet, especially on social media platforms such as Twitter and Facebook. NER operates as a component of information extraction, intending to identify and classify specific elements of text into predefined categories, such as the names of individuals, organizational entities, and geographic locations, an essential process for this research. The NER models utilized in this research are detailed as follows:

4.3.1 SpanMarker

SpanMarker introduces a novel approach to addressing tasks related to NER [25]. Instead of defining NER as a typical sequence labeling task, SpanMarker conceptualizes it in terms of spans. Initially, SpanMarker identifies potential span "phrases" before determining if these spans represent entities and, if so, what specific type of entity they correspond to. The SpanMarker model excels particularly in fine-grained NER applications, where precision is paramount to distinguishing entities across various types and classes within specialized vocabularies. Trained on the FewNERD dataset, the model recognizes a wide range of specific and detailed entity categories, such as distinct types of individuals, organizations, or locations [25]. Users can customize various aspects of the model, such as the input size, span lengths for markers (in both directions), and maximum length allowed for entities, through parameters like "model_max_length" and "en-

tity_max_length," offering comprehensive control over input characteristics, span dimensions, and entity sizes, rendering the model highly adaptable. The model's implementation structure in this research is shown in Chapter 6.

4.3.2 spaCy

spaCy is an open-source Python library focused on advanced NLP. It refines various pre-trained models tailored for different languages and tasks, encompassing tokenization, part-of-speech tagging (POS), dependency parsing, NER, and more [26]. These models prioritize speed, accuracy, and user-friendliness, making them popular for research and production. While some spaCy models are pre-trained, others are not, offering a spectrum of sizes and complexities to suit diverse needs regarding speed and precision. The models in spaCy are typically classified as small (sm), light and fast with slightly lower accuracy; medium (md), featuring a balance of size and accuracy; and large (lg), optimized for tasks like named entity recognition or parsing, albeit at a slower pace and higher resource consumption [26]. Despite being pre-trained, spaCy models can be fine-tuned for improved performance in specific domains using custom data, which is beneficial in scenarios where default models fall short. This flexibility to fine-tune models serves well for research and production purposes, offering a range of NLP solutions from efficient small models to highly accurate large ones. SpaCy's NER model was employed in conjunction with the SpanMarker model, as detailed in Chapter 6.

4.4 Data Extraction Models

Data extraction involves retrieving and gathering specific data from diverse sources, such as databases, websites, and documents, in structured and unstructured formats. This process is crucial in data analysis, business intelligence, and research. Various techniques like web scraping, database queries, and file parsing are used for data extraction, often serving as a key step in the ETL (Extract, Transform, Load) process. Data is extracted, then transformed to enhance usability, and finally loaded into a system for viewing or further processing. This research uses the PDFplumber model to extract structured data from historical travelogues written in 19th-century English [Chapter 6].

PDFplumber is a Python library designed to extract structured data from PDF files. It includes tools for separating text, tables, images, and metadata from PDFs while maintaining the layout structure. Its workflow is especially suited to extracting content from complex documents with two-column layouts, embedded fonts, or dense tables. Additionally, PDFplumber provides precise control over individual page elements, making it valuable for applications that demand accurate data extraction from PDFs. Furthermore, it incorporates Optical Character Recognition (OCR) corrections and can be integrated with tools like Tesseract to process scanned PDFs containing text as images rather than selectable text. By utilizing OCR to extract text, PDFplumber offers features

to enhance the accuracy and alignment of the extracted content with the original document's layout. It proves beneficial when working with poorly scanned or low-quality PDFs, enabling users to adjust and refine the OCR output as needed.

4.5 Large Language Models (LLM)

Large Language Models (LLM)s are transformer-based models utilizing deep learning to process and understand natural language [27]. They are trained on extensive text data to recognize patterns in unstructured sequences and acquire a language knowledge base. One key advantage of LLMs over traditional NLP models is their ability to perform various tasks effectively with minimal input. They can undertake tasks they were not explicitly trained in using human language descriptions or task examples. This efficiency translates to requiring fewer resources and less training time while maintaining comparable performance to conventional models. LLMs excel in tasks like text generation, translation, and summarization, with applications in virtual assistants, chatbots, and language translation systems. Prominent LLMs include BERT by Google [28], GPT by OpenAI [29], and LLaMa by Meta [30].

The transformer model represents a significant advancement in natural language processing (NLP), leveraging an attention mechanism to process long input text sequences through its encoder-decoder architecture efficiently [31, p. 2]. The encoder block is responsible for generating numerical representations, or embeddings, of the input text. In contrast, the decoder block interprets these embeddings to produce coherent and contextually relevant output text sequences. This architecture has become the foundation for many state-of-the-art language models, including ChatGPT.

ChatGPT, a leading example of a large language model (LLM), is built upon the InstructGPT framework, which introduced a groundbreaking methodology to integrate human feedback into the training process of transformer-based models. This approach, called Reinforcement Learning from Human Feedback (RLHF), aligns the model's outputs more closely with user intentions and ethical guidelines. OpenAI's seminal 2022 paper, *"Training Language Models to Follow Instructions with Human Feedback"*, provides a comprehensive exploration of RLHF, detailing how human evaluators guide the model's learning process to enhance its performance and alignment with user goals [32]. This process is visually summarized in Figure 4.3, illustrating the iterative feedback loop between human input and model refinement. By combining the transformer's robust architecture with RLHF, ChatGPT exemplifies how advanced NLP systems can achieve remarkable adaptability and user-centric performance.

This research leverages LLM models such as GPT by OpenAI [29], LLaMa by Meta [30], and Mistral by Mistral AI [33] to implement annotation guidelines and perform automated annotation. Utilizing LLMs ensures consistency and accuracy in the annotation process, which is crucial for developing robust NLP systems. The implementation process for LLM models is detailed in Chapter 7.

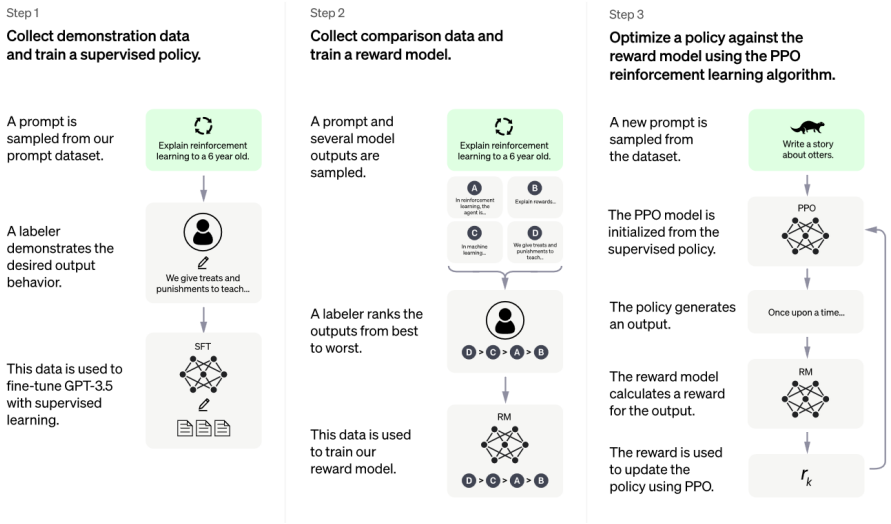

Step 1
Collect demonstration data and train a supervised policy.
A prompt is sampled from our prompt dataset.
Explain reinforcement learning to a 6 year old.
A labeler demonstrates the desired output behavior.
We give treats and punishments to teach...
This data is used to fine-tune GPT-3.5 with supervised learning.
SFT
Step 2
Collect comparison data and train a reward model.
A prompt and several model outputs are sampled.
Explain reinforcement learning to a 6 year old.
A labeler ranks the outputs from best to worst.
D > C > A > B
This data is used to train our reward model.
RM
D > C > A > B
Step 3
Optimize a policy against the reward model using the PPO reinforcement learning algorithm.
A new prompt is sampled from the dataset.
Write a story about otters.
The PPO model is initialized from the supervised policy.
PPO
The policy generates an output.
Once upon a time...
The reward model calculates a reward for the output.
RM
The reward is used to update the policy using PPO.
r_k

Figure 4.3: A diagram illustrating the three steps of instructGPT: (1) supervised fine-tuning (SFT), (2) reward model (RM) training, and (3) reinforcement learning via proximal policy optimization (PPO) on this reward model. [32, p 3]

.

Prompt Engineering

Prompt engineering involves crafting prompts or instructions that direct a language model's output, such as ChatGPT [34, p 6]. It empowers users to influence the model's responses and produce text customized to their specific requirements. A cutting-edge language model excels at generating text resembling human writing. Leveraging the transformer architecture, it can manage vast volumes of data to generate top-notch text. To optimize LLM 's performance, mastering how to prompt the model correctly is crucial. Prompting facilitates users in regulating the model's output to produce pertinent, precise, and high-quality text. Understanding LLMs proficiencies and constraints is vital when interacting with the model. While adept at mimicking human text, the model may fall short of desired outcomes without appropriate guidance. This is where prompt engineering plays a pivotal role; by offering precise and detailed directives, users can steer the model's output to ensure relevance. A prompt formula consists of three key components as follows [34, p 6]:

- **Task** - a clear and succinct statement requesting the model's generated output
- **Instructions** - guiding the model in text generation and
- **Role** - defining the model's stance in text creation.

Various in-context learning techniques can be employed to generate text from LLMs with minimal or no examples. These approaches prove beneficial when data are scarce for a specific task or when

the task itself is novel and vaguely defined. The techniques are described as follows [34, p 14]:

Zero Shot Prompting

The zero-shot prompting technique is used when no examples are accessible for the given task. In this method, the model is tasked with a general prompt and then generates text based on its comprehension of the task.

One Shot Prompting

When only one example is available for the task, the one-shot prompting technique is utilized. The model is presented with the sole example and produces text based on its interpretation of that example.

Few Shot Prompting

For scenarios with a limited number of examples for the task. The model is furnished with these examples and generates text based on its understanding.

To implement the proposed annotation guideline, prompt engineering plays a central role, employing two-shot and few-shot prompting strategies to evaluate performance in this context [Chapter 7].

4.6 Evaluation Matrics

Precision, recall, and F1 score are fundamental metrics for evaluating the performance of the proposed classification model, each providing unique insights into the model's effectiveness [35]:

Precision: This metric represents the proportion of true positive predictions among all positive predictions made by the model. It quantifies the model's ability to avoid incorrectly labeling negative instances as positive, thereby reflecting its accuracy in identifying relevant cases. High precision indicates a low rate of false positives.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall: Recall measures the proportion of true positive predictions among all actual positive instances in the dataset. It evaluates the model's ability to identify all relevant instances, ensuring that positive cases are not overlooked. High recall indicates a low rate of false negatives.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both measures. It is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is particularly valuable for imbalanced datasets or scenarios where the costs of false positives and false negatives differ significantly. By considering both precision and recall simultaneously, the F1 score comprehensively evaluates the model's performance, ensuring that neither metric is disproportionately prioritized.

These metrics collectively provide a robust framework for assessing classification models. This enables researchers to identify strengths and weaknesses in model performance and optimize their approaches accordingly.

Chapter 5

Annotation Guidelines

This chapter presents the proposed annotation guidelines and explicitly describes the annotation classes with sample examples prepared during an 18 credit research-oriented project [36].

5.1 Task Objective, Details and Setup

The goal is to determine and describe the connection or relationship between two specific characters or a specific character (Traveler) and a designated place/ location within a given excerpt. This entails identifying the nature of their association and providing an explanation or description of it (This will help us to find valid itineraries of a traveler or group of travelers.). [3]

e.g.: **John Stewart**, Esq. was born in **London** in the year 1749 [18].
e.g.: **I** (WILLIAM MACMICHAEL) shall always reflect with great satisfaction on the adventures of a journey during which I enjoyed the pleasure of the society of my friend **Mr. Legh** [19].

What is a character?

A character refers to a human or a person entity within a narrative that possesses the capability to engage in monologues or dialogues. An automated entity tagger is employed to identify characters, with entities categorized as PER (person) [Section 4.2] according to the guidelines specified in a particular research paper, being regarded as potential characters. [3]

e.g.: **Mr. Stewart** displayed the most fanciful and yet pleasing taste [18].

What is a place?

A place refers to a defined spatial area that can be physically demarcated. It encompasses locations such as rooms, cities, or specific geographical features. Similar to character identification, candidates for place entities are automatically recognized using an entity tagger. In this case, entities falling under the categories of LOC (location), FAC (facility), or GPE (geopolitical entity) [1]

[1]Note: All examples used in this chapter are in the public domain, eliminating potential copyright infringements.

[Section 4.2], as outlined in a specific research paper, are regarded as potential places. [3]

e.g.: Mr. Stewart returned to **England** (GPE) [18].

Task details and setup

The excerpts will be provided from the source of historical travelogues of the Ottoman Empire (In English language) [1]. Each excerpt will have a highlighted section representing a character or characters and a separate highlighted portion representing a place. Identifying these character and place spans will be achieved through an automated system, eliminating the need for manual intervention in the process. If an annotator finds a place not correctly recognized by the tagger, then the annotator will provide the label themselves.

5.2 Annotation Classes

Each excerpt is assigned seven distinct annotations to capture various travel aspects [Table 4.1].

Table 5.1: Annotation Classes

Annotation	Type (Domain)	Possible values
Valid relation [3]	Binary (PER-LOC)	• TRUE • FALSE
Spatial relation [3]	Categorical (PER-LOC)	• IN • NEAR • THRU (Through) • TO • FROM • NO REL
Temporal span [3]	Binary (PER-LOC)	• PUNCTUAL • HABITUAL
IsPartOfJourney	Binary (PER-LOC)	• TRUE • FALSE
Valid connection	Binary (PER-LOC)	• TRUE • FALSE
traveling relation/ Travel relationship / Travel connection	Categorical (PER-PER)	• IN GROUP • JOIN GROUP • EXIT GROUP • NO REL

The annotation classes will be described in more detail next, but before that let's preview the annotations for a simple example. Consider the following sentence: "**<PER id=0>**Alex**</PER>** and **<PER id=1>**Maya**</PER>** **<Action>**started**</Action>** their adventurous journey **<Action>**from**</Action>** the mountainside village called **<PLACE Itinerary_ID=1>**Hallstatt**</PLACE>**."

For annotation purposes, `'id=0'` has been assigned to the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and `'id=1'` to another character. Additionally, the keywords that refer to spatial relationships have been labelled as `'Action'`. Furthermore, an `'Itinerary_ID'` number is assigned to the `<PLACE>` where `'IsPartOfJourney=TRUE'`, in a way that represents the primary character's actual movement. Here are the annotations for the characters *Alex, Maya, and the place "Hallstatt"*:

- **Valid Relation:** TRUE (Both characters and the place are correctly tagged.)
- **Spatial Relation:** FROM (Alex and Maya started the journey "From" the Hallstatt, indicating their departure from the place.)
- **Temporal Span:** PUNCTUAL (The journey is presented as a single, specific event.)
- **IsPartOfJourney:** TRUE (The narrative indicates the start of a journey, and they started from 'Hallstatt' together)
- **Valid connection:** TRUE (The narrative indicating valid connection between the primary character and the other character.)
- **Travel connection:** IN GROUP (The narrative describes Alex and Maya traveling together, which indicates a valid travel connection between the characters)

This annotation setup ensures that the primary character's movement is accurately represented, with the spatial relationship clearly indicated through the labelled `'Action'` and the place assigned an `'Itinerary_ID'`. The place *'Hallstatt'* is part of the journey (IsPartOfJourney=TRUE) and is appropriately annotated to reflect its significance in the itinerary.

Annotation Class Descriptions and Examples

The annotation classes are described in more detail below, along with examples to illustrate their application:

Valid Relation

Not all references to people and places indicate entities that can be directly connected to each other; a significant group that doesn't fall into this category are generic references, as explained by Reiter and Frank (2010) [17], shown in the following example:

e.g.: He then struck through the country to the westward [18].

In this context, the country or the westward denotes a generic type of entity, not a specific place within the story's world. To streamline the process, the instances are eliminated where the focal character or location cannot be linked, whether due to their generic nature or to inaccuracies in entity identification. This is formalized as a binary task, where an illustration is annotated "TRUE" if accurately identified by the entity tagger, where both the character and place can be connected; If the character or the place is not correctly identified, the annotation should be "FALSE" [3].

Furthermore, the initial annotation serves as a filtration process, aimed at extracting a subset of high-quality instances devoid of entity tagging discrepancies. Moreover, it discerns generic characters and locales as undesirable entities. To clarify these connections, the illustrative instances are provided as follows[1]:

- **FALSE**: In cases where the designated place or character is inaccurately labelled or the highlighted text refers to a non-specific place or non-specific character (or both). For example-

  1. "When a man of genius, remarkable for talent or noted for eccentricity, leaves this <PLACE>world</PLACE> for ever and ever, the <PER>public</PER> mind immediately seeks knowledge of his private manners, and eagerly receives every information the press may afford, provided such communication be true and authentic." [18]

     Explanation: In this context, the identified place and person, *"world"* and *"public"* are far too generic. The identified place and person should be pinpointable, specific, and precise.

  2. "<PER>I</PER> am in hope it will not be construed into presumption to solicit the clemency of those merciless rogues the <PLACE>Edinburgh</PLACE> <PER>Reviewers</PER>, who with that acrimony so peculiar to critics who have the false idea that their profession necessarily compels them, butcher-like, to cut up." [18]

     Explanation: In this context, the place is incorrectly tagged, *"Edinburgh"*, is referring to a group of people or organisations ("Edinburgh Reviewers") instead of a place. On the other hand, *"Reviewers"* is far too generic as a person entity, like example 1.

  3. "But she looked at me once more, as if she believed I was still curious and might inquire <PER>someone else</PER>. Then she mentioned she had heard <PLACE>it</PLACE> was once a secluded monastery."

     Explanation: The reference to *"someone else"* and *"it"* are a non-specific person and place entity.

  4. "However, <PER>they</PER> do at times some good to us poor <PER>authors</PER>." [18]

     Explanation: Similar to (1 and 2), *"authors"* are far too generic as a person. Mentions of

groups of individuals (e.g., <PER>Reviewers</PER> are frequently indolent) should be classified as "FALSE" due to their lack of specificity. These references do not pertain to identifiable individuals but rather describe a general category of people.

5. "For decades after the division, when the region had established its own identity and culture, the northern territories were still remembered by <PER>many</PER> as the <PLACE>homeland</PLACE>."

   Explanation: In this context, the candidate term *"millions"* does not refer to an individual or a specific group, making it unsuitable for annotation as a named entity. Similarly, the term *'homeland'* is too generic to qualify as a specific place entity, as it lacks the precision required for meaningful entity recognition.

6. "The silence was absolute; not a whisper from the gentlest <PER>zephyr</PER> could be heard; the <PLACE>lake</PLACE> before them, which was unusually calm, appeared frozen in time; not a single branch swayed; the tall pines stood as motionless as statues; and the atmosphere was thick with the scent of pine and earth."

   Explanation: In this context, *"zephyr"* means a soft breeze; hence *"zephyr"* is incorrectly tagged by the entity tagger with the person tag.

- **TRUE**: When both the character and place are accurately labelled, and they pertain to specific, non-generic character and place references, respectively. For example-

  1. "On Ash Wednesday morn, this good and amiable man died, without a pang, at <PER>his</PER> residence in <PLACE>Northumberland-street</PLACE>." [18]

     Explanation: In this context, both *"his"* and *"Northumberland-street"* are correctly tagged and non-generic. Pronominal references (e.g., he, she, they, it) and adverbial references (e.g., here, there) should be eligible for annotation, even when the specific named entity being referred to is not immediately evident from the linguistic context. If these pronouns can be successfully linked to a valid, non-generic named entity, they should be considered valid spans.

  2. "The wind had been very high while <PER>we</PER> were in the <PLACE>port of Falmouth</PLACE>, and the weather was still unsettled." [1]

     Explanation: The span for the place is correctly identified and specific.

  3. "<PER>He</PER> therefore went to <PLACE>Margate</PLACE> to inhale the refreshing breezes of the sea; and returned back to <PLACE>London</PLACE> little the better for this trip." [18]

     Explanation: In this context, *"He"*, *"Margate"* and *"London"* are correctly tagged and non-generic.

4. "The Philosopher, however, was somewhat aroused from the calm and tranquil life <PER>he</PER> passed in <PLACE>Cockspur-street</PLACE>, by the sudden arrival of <PER>Her late Majesty, Queen Caroline</PER>, whose appearance in <PLACE>England</PLACE>, was the signal for riot and tumult among the lower classes of the population." [18]

   Explanation: Similar to the above examples, both the person and place entities are correctly tagged and non-generic.

5. "<PER>He</PER> therefore, doubtful of the course of things, resolved to leave the country for <PLACE>America</PLACE>, where he hoped to meet with a refuge from the impending storm." [18]

   Explanation: In this context, *"He"* and *"America"* are correctly tagged and non-generic.

6. "<PER>He</PER> then struck through the country to the westward, and fortunately, before the alarm had sufficiently spread, succeeded in effecting his escape to one of the <PLACE>European forts</PLACE>, where, by the skill of a surgeon in the Company's service, his wound was cured." [18]

   Explanation: Even though *"European forts"* does not seem like a specific place, in the context it's clear that it's specific to the place where the character is currently on. The presence of multiple potential locations for a place does not inherently categorize it as generic. For instance, *"Cambridge"* is found in both the United States and England, and the linguistic context may not provide any clues regarding which specific place is being referred to. Despite multiple possibilities, the place should not be considered generic.

Furthermore, the personification of non-human entities (e.g., a talking dog) should not be labelled as BAD CHAR, provided that the description clearly indicates the entity possesses cognitive abilities or characteristics akin to a human.

Spatial relation

This study employs a categorical annotation scheme to capture the spatial relationship between a character and a place, following the guidelines adapted from Soni et al. [3]. These annotations are designed to provide structured and interpretable representations of spatial interactions within the text, enhancing contextual understanding of relationships between entities. By incorporating these guidelines, the annotation process ensures consistency and aligns with established methodologies for grounding spatial relationships in textual data.

This study assigns spatial relationships based on the primary character's interaction with the place entity to determine an itinerary within the given context. In total, there are six spatial relationship categories, as outlined in Table 5.1, with their definitions provided in Table 5.2 [3].

Table 5.2: Short description of spatial relation's subtask

Category	Short description
IN	The character (C) is at, in, or on the place (P)
NEAR	The character (C) is proximity closed to the place (P) but not at the place (P)
THRU	The character (C) is passing by/through the place (P)
TO	The character (C) is going to/ moving toward's the place (P) and is certain to reach the place (P)
FROM	The character (C) was at the place (P) and has moved from the place (P)
NO REL	There is no spatial relationship between the character (C) and the place (P)

The primary annotation pertains to the spatial association between the primary character reference and a place reference. The keywords denoting spatial relationships have also been labelled as "`Action`". These, as well as the following annotations, should only be applied when the valid relationship has been designated as "`TRUE`". Various values can be assigned to this annotation, to clarify these categories, the illustrative instances are provided as follows[1]:

- **IN:** The category indicates whether a primary character is within the spatial boundaries that define the location. This association can be explicitly indicated by prepositions like at, in, or on, or it may need to be deduced. The key determination to be made is whether the primary character's position falls within the geographical confines of the place. While lexical indicators can aid in this determination when available, this relationship may need to be inferred from the context when such cues are absent. To clarify these connections, the illustrative instances are provided as follows:

  1. "<PER>He</PER> had fondly cherished an idea that his doctrines would have been received with enthusiasm by the English; and having long digested that theory, he was prepared, when <Action>settled in</Action> <PLACE>London</PLACE>, to make it public." [18]

     Explanation: Because the character referred to by *"He"* (possibly the narrator) indicates he would be settled in *"London"*, it can be inferred that the character is "`IN`" the place.

  2. "The wind had been very high while <PER>we</PER> were <Action>in</Action> the <PLACE>port of Falmouth</PLACE>, and the weather was still unsettled." [1]

     Explanation: It is evident in this excerpt that the marked characters were at *"the port of Falmouth"* during the narrative period.

  3. "<PER>Our</PER> <Action> entry into </Action> the <PLACE>Mediterranean</PLACE> is here faintly described, as no words can convey the ideas, excited by scenes of so much novelty, grandeur, and beauty." [1]

Explanation: The characters, referred to by *"Our"* is entering into *"the Mediterranean"*, indicating their presence in that location.

4. "<PER>We</PER> <Action>stood for</Action> <PLACE>Corsica</PLACE> with a brisk gale and a great swell, which took us on the weather side; the waves distinct, vast, and black, breaking with white tops." [1]

   Explanation: The character, referred to by *"We"* is stood for *"Corsica"*; hence, the character is in the place.

5. "<PER>We</PER> <Action>tarried at</Action> <PLACE>Genoa</PLACE> until the 25th of July." [1]

   Explanation: The category falls under "IN" because the motion is completely contained within the boundaries of the place.

- **NEAR:** This label indicates if a primary character is near a location but not inside it. This assessment is largely based on context and perspective; just as discourse can adjust the similarity between two entities for co-reference purposes [37], conversation can also alter the perceived closeness of a character to a place.

  To determine proximity, the distance between a character and a celestial body can be vast yet still considered near, while being in one city like Goettingen may be far from another, such as Paderborn. Evaluating the size of the location is recommended when drawing a boundary to define what constitutes "near". By default, a bedroom would be deemed close to a living room but not to a city hall, illustrating the influence of scale on proximity judgments. *For example, Berlin could be close to Potsdam but distant from Bangladesh (or Mars).* Therefore, size acts as a proxy for categorizing the similarity of places and should inform assessments of closeness. Instances may arise where apparent short distances are actually considered extensive, as illustrated in statements like *"Feeling exhausted, she knew there was still another 500 feet to go before she reached the summit"* [3]. In such instances, they should be classified as "NO REL" (see "NO REL" 5.2). To clarify this category, the illustrative instances are provided as follows:

  1. "Our attention had been, for some time, agreeably engaged by the classical country, which surrounded us, and <PER>we</PER> were now <Action>near</Action> <PLACE>Sunium or Cape Colonne</PLACE>, and <PLACE>the coast of Attica</PLACE>." [1]

     Explanation: It's clear in this excerpt that the character is near the places.

  2. "As <PER>we</PER> now <Action>approached near to</Action> the <PLACE>Mediterranean</PLACE>, some of the sailors had got a strong new rope, and prepared it for ducking such of the crew as were novices in this sea." [1]

Explanation: It's not clear in this excerpt whether the character is in *"Mediterranean"* but it's certain that the character is near to the place (since the character is approaching near to the place). Hence, the label is "NEAR".

3. "The Pharos of <PLACE>Genoa</PLACE> <Action>appeared</Action> as a tall pillar, the coast picturesque and mountainous, its slopes." [1]

   Explanation: It's unclear whether the character entered *"Genoa"* but the character surely is in the nearest place (again, the character can see it from their position). This example highlights two important things: (i) The boundaries of the place may not be immediately clear and have to be judged; (ii) nearness is a matter of interpretation from the annotator.

4. "<Action>Not far</Action> from thence is <PLACE>Chelindreh</PLACE>, a snug but very small port." [38]

   Explanation: The narrator seems close to *"Chelindreh"* (*"Not far"* means shortest distance, which can assume as near). This is why the annotation is as "NEAR".

5. "<PER>I</PER> <Action>passed</Action> the morning of December 25th, <Action>opposite</Action> the village of <PLACE>Massoura</PLACE>." [39].

   Explanation: It is unclear in this excerpt whether the character is in *"Massoura"*. However, the character is inevitably in the opposite direction of *"Massoura"* (since he passes the morning opposite of *"Massoura,"* which indicates the character is near the place). Hence, the label is "NEAR".

- **THRU:** Unlike prior categories where primary character is often stationary, this category suggests movement through a location. "THRU" typically conveys an "IN" relationship as well, but offers a more detailed perspective on the characteristics of that relationship. Importantly, this category stipulates that the character's starting point and ultimate destination are distinct from the place they are identified as traversing. Furthermore, this applies when the character starts at place X (which is not Y), traverses through place Y, and concludes at place Z (which is not Y). The character should not have originated or concluded their journey at place Y but should solely pass through it. While this category implicitly suggests that the character is within the place, the pivotal differentiation for "THRU" is that the character is presumed to be in motion, whereas "IN" may or may not involve movement. To clarify this category, the illustrative instances are provided as follows:

  1. "<PER>Mr. Stewart</PER> <Action>took leave of</Action> <PLACE>India</PLACE>, and <Action>traveled over</Action> near the whole of <PLACE>Persia</PLACE> and <PLACE>Turkey</PLACE> on foot." [18]

     Explanation: The character *"Mr. Stewart"* walks through *"Persia"* and *"Turkey"* (implying that he started his journey outside the places and also ended outside of those places).

Thus, the annotation is "THRU".

2. "<PER>He</PER> <Action>made the tour of</Action> <PLACE>Scotland</PLACE>, <PLACE>Germany</PLACE>, <PLACE>Italy</PLACE>, and <PLACE>France</PLACE>, on foot, and ultimately <Action>settled in</Action> <PLACE>Paris</PLACE>." [18]

   Explanation: The action tag *"made the tour of"* indicates that the places fit the definition of passing through places; the action starts outside of the mentioned places, carries the character through it, and ends when he has reached *"Paris"* (at the end of the action, the character is no longer in those places).

3. "A gentle breeze springing up, on the evening of the 14th of August, <Action>wafted</Action> <PER>us</PER> by <PLACE> Maritimo, a rocky island </PLACE>, on which is a fishing town." [1]

   Explanation: In this context, the text implies motion through the marked place before reaching a different destination. Hence, "THRU".

4. "On the second day <PER>we</PER> <Action>passed</Action> the <PLACE>island Gorgogna</PLACE>, by which were many sail of small-craft fishing under shore for anchovies." [1]

   Explanation: The marked place (*island Gorgogna*) is an intermediate place, the characters are moving towards a different destination. This is the definition of "THRU".

5. "<PER>We</PER> <Action>sailed by</Action> <PLACE>Cape Malea, now St. Angelo</PLACE>, the sea almost smooth, but the waves swelling at intervals, with a hollow noise, and seeming to pursue us." [1]

   Explanation: Similar to example- 3, the action tag implies motion through the marked place before reaching a different destination. Thus, the annotation is "THRU".

- **TO:** This classification delineates a primary character moving towards a specific destination, implying that the destination has undeniably been reached. Similar to the "THRU" category, this classification typically involves a connection with the destination, while offering a more detailed description of the movement.

  Furthermore, if the character is progressing towards the place mentioned in the example, and it is either explicitly stated in the text or reasonably inferred that the character has unquestionably reached the place. The distinction between "TO" and "IN" lies in the fact that the text clearly describes how the character reached the place in the former. At the outset of the action, it is inferred that the character is not yet within Y, but by the end, it's apparent that they have arrived there. If there is any doubt regarding whether the character reached the destination, the annotation should be categorized as "NO REL" (see "NO REL" 5.2). To

clarify this category, the illustrative instances are provided as follows:

1. "<PER>He</PER> accordingly took a <Action>passage on</Action> board a ship <Action>bound to</Action> <PLACE>New York</PLACE>, and paid the Captain the sum required." [18]

   Explanation: The sentence describes the character embarking on a journey from one place to another. Additionally, the mention of the ship being bound for "*New York*" provides explicit confirmation that the character will reach the destination, justifying the label as "TO".

2. "In the morning <PER>we</PER> <Action>approached</Action> <PLACE> Cythera or Cerigo</PLACE>." [1].

   Explanation: The characters approached "*Cythera or Cerigo*" but from the context available here, it's not clear if "*The characters*" reached "*Cythera or Cerigo*". It is indicated with certainty, though, that "*The characters*" left their previous place on the way towards "*Cythera or Cerigo*".

3. "<PER>He</PER> therefore <Action>went to</Action> <PLACE>Margate </PLACE> to inhale the refreshing breezes of the sea." [18]

   Explanation: In this context, it's clear that the character moved towards and got to the "*Margate*".

4. "The next day <PER>we</PER> weighed anchor again, and <Action>proceeded to </Action> <PLACE>Falmouth</PLACE> to complete our cargo." [1]

   Explanation: It's clear that the characters proceeded to the place marked and eventually reached the place.

5. "<PER>He</PER> determined on acquiring the languages of the country, and to become an interpreter, until by the exercise of economy <PER>he</PER> had saved £3000, then to purchase an annuity and <Action>return to</Action> <PLACE> England</PLACE>." [18]

   Explanation: The character referenced by "*he*" returns to "*England*", indicating his movement towards it.

- **FROM:** This category denotes motion, wherein a primary character previously occupied a location and has since relocated from it or when the primary character had been present at the specified place but has subsequently departed from it to transition to a new location. The "FROM" label is contingent on both the time and the designated place. Key considerations include: -Apply this label when the character's last static position is at the specified place. -If the text indicates the time elapsed since being at the marked place, this should be a factor in the annotation. For example, "*left the kitchen*" should be handled differently than "*left*

the country". If a significantly large amount of time has passed in the former case (e.g., a month), assigning the "NO REL" label may be more appropriate. To clarify this category, the illustrative instances are provided as follows:

1. "Many years after <PER>he</PER> had <Action>quitted</Action> <PLACE> India</PLACE>, and as totally unexpected by him at the time the fortunate news arrived." [18]

   Explanation: Here "*he*" is the character and with respect to the marked place, it's explicitly stated that the marked character went away from "*India*".

2. "We may nevertheless attribute <PER>Mr. Stewart's</PER> prolongation of his intended trip, to the advice and persuasions of his faithful friends, whose remonstrances were happily attended with effect, otherwise <PER>he</PER> certainly had taken <Action>leave of</Action> <PLACE>England</PLACE>." [18]

   Explanation: The "*he*" refers to "*Mr. Stewart*". It's said that he took leave of "*England*", the marked place. Because of the word "*leave*", it can be inferred that "*Mr. Stewart*" was in "*England*" before but is not there anymore.

3. "<PER>He</PER> took his <Action>departure from</Action> <PLACE>Paris </PLACE> <Action>to</Action> <PLACE>England</PLACE>." [18]

   Explanation: The sentence describes that the character started his journey from one place to reach another place. Moreover, the character moved from "*Paris*", so the label is "FROM".

4. "<PER>We</PER> were <Action>becalmed</Action>, on the 17th of July, <Action>off</Action> <PLACE>Cape de Melle</PLACE>; and then had a fine gale, and <Action>approached</Action>, <PLACE>Italy</PLACE> at the rate of twelve miles an hour." [1]

   Explanation: In this context, it's clear that the characters started their journey from "*Cape de Melle*" towards "*Italy*". So the label is "FROM".

5. "<PER>We</PER> weighed anchor and <Action>got out of</Action> the <PLACE>mole</PLACE> in the night." [1]

   Explanation: Here "*We*" the characters and with respect to the marked place, it's explicitly stated that the marked characters went away from "*mole*".

- **NO REL:** Identifying primary character-place pairs that occur together in the text without explicitly implying a spatial relationship between them is one of the complicated tasks among other spatial relationships. The term "NO REL" indicates the absence of such a connection, compared to scenarios where there is insufficient information to determine the link between the character (C) and the place (P). The concept of "NO REL" can be applied when a character

is in the process of moving towards a destination. Yet, it remains uncertain whether the destination has been reached or epistemic narrative ambiguity exists in the text—where the narrator or characters are unsure of the relationship between the character and place. As demonstrated in the examples above, the relationships between a character and a place are frequently subtly implied, heavily relying on the reader's inference rather than explicit spatial cues within the text.

Furthermore, when none of the previously described spatial relationships apply, the appropriate label is "NO REL". The "NO REL" label applies in various scenarios, including but not limited to [3]- (i) When the text unambiguously indicates no discernible relationship between the character and the place. (ii) When insufficient information is available in the text to judge the spatial relationship between the character and the place. (iii) When the spatial relationship implied in the text is speculative or hypothetical. (iv) When the character is in the process of moving toward the place, but it remains unclear if they ever reach the destination. (v) When it is definitively established that the character is not within the place. (vi) When the text explicitly conveys uncertainty regarding the character's connection with the place. (vii) When the text creates an illusion that shorter distances are, in fact, longer, the "NEAR" label should not be applied, and "NO REL" is the more suitable choice. To clarify this category, the illustrative instances are provided as follows:

1. "<PER>John Stewart</PER>, Esq. was <Action>born</Action> in <PLACE> London</PLACE> in the year 1749." [18]

   Explanation: In this context, the action tag "*born*" does not indicate any movement nor does it fall into any other spatial category. So the label is "NO REL".

2. "The <PLACE>Theatre of Herodes Atticus</PLACE>, was <Action>built by</Action> <PER>Herodes Atticus</PER> in honour of his deceased wife <PER>Regula</PER>." [40]

   Explanation: Similar to example- 1, the narrative does not fall into any other spatial category. So the label is "NO REL".

3. "<PER>I</PER> formed was <Action>to go by land to</Action> <PLACE>Aiolou Bourgas</PLACE>" [20]

   Explanation: The spatial relationship between the character and the place will possibly activate in the future. Because the current relationship is hypothetical, the label is "NO REL".

4. "<PER>I</PER> <Action>shall continue my journey to</Action> that city, and thence <Action>proceed to</Action> <PLACE>Vienna</PLACE>." [20]

   Explanation: This sentence has two places – "*Vienna*" and "*that city*". The sentence

implies that "*that city*" differs from "*Vienna*", the marked place. The character, "*I*", is moving toward "*that city*" which means in the current scenario "*I*" is not in or moving toward "*Vienna*". Thus, this example is marked with "*NO REL*".

5. "<PER>We</PER> were becalmed, on the 17th of July, <Action>off</Action> <PLACE>Cape de Melle</PLACE>; and then had a fine gale, and <Action>approached</Action> <PLACE>Italy</PLACE> at the rate of twelve miles an hour. " [1].

   Explanation: It's clear that the character is in motion towards the destination, the marked place, but it's unclear from the limited text whether the character reached the place "*Italy*".

In instances of ambiguity, the order of precedence for the labels should be as follows: [TO, THRU]> IN > [NEAR, FROM] > NO REL [3]. This prioritization places the categories that suggest character movement before the categories in which the character remains stationary. To illustrate this concept further, let's consider the following examples in Table 5.3:

Table 5.3: Examples of ambiguous instances

Example	**Possible categories**	**Preferred category**
"<PER>He</PER> was <Action> walking through </Action> <PLACE> the forest </PLACE>." [3]	**IN**: if it's equally likely that the character is contained within the place **THRU**: if it's equally likely that the character started outside the forest, and ended up outside the forest but was transiting through the forest	THRU (Through)
"<PER>They</PER> continued to pace themselves <Action> to reach</Action> <PLACE> the summit</PLACE> before noon." [3]	**TO**: if it's equally likely that the character is moving towards the destination and reached there **NO REL**: if it's equally likely that despite the motion towards the destination the character may not have reached the destination	TO
<PER> The team </PER> was exhausted <Action> to continue the last mile towards </Action> <PLACE> the camp </PLACE>." [3]	**NEAR**: if it's equally likely that the character is close to the place marked **NO REL**: if it's equally likely that the perceptive distance is larger	NEAR

The following subtasks take into account the duration of the spatial relationship (e.g. Temporal span) and the present status of the relationship in relation to the narrative timeline.

Temporal span

Characters may establish either a short-term or long-term spatial connection with a location. When determining whether a primary character is situated in their primary residence or city, it is essential to distinguish between their physical presence at a specific moment and a habitual association with that place, which may not indicate immediate presence at that instance. This distinction is designated by indicating the duration of the spatial relationship. The duration of the relationship indicates the length of time a character spends in that particular place. If the connection is short-term or momentary, it is termed *punctual*, as demonstrated in the following instance:

"<PER> Mr. Stewart </PER> <Action> traveled over </Action> near the whole of <PLACE> Persia </PLACE> and <PLACE> Turkey </PLACE> on foot." [18]

Conversely, if the link is long-term, seasonal, or recurrent, it is regarded as *habitual*, as illustrated below:

"He had fondly cherished an idea that his doctrines would have been received with enthusiasm by the English; and having long digested that theory, <PER> he </PER> was prepared, when <Action> settled in </Action> <PLACE> London </PLACE>, to make it public." [18]

In this depiction, although it is not explicitly mentioned how long the specific character has stayed in London at the time of narration, one can infer the character possesses a habitual affiliation with the city as he has settled in London. To illustrate this concept further, let's consider the following examples[1]:

- **PUNCTUAL:** If the primary character and the place have a transient spatial connection, it should be labelled as "PUNCTUAL". For example:

  1. "<PER> He </PER> <Action> made the tour of </Action> <PLACE> Scotland </PLACE>, <PLACE> Germany </PLACE>, <PLACE> Italy </PLACE>, and <PLACE> France </PLACE>, on foot." [18]

     Explanation: From the limited context available, the character seems to have a one-time relationship with the place.

  2. "<PER> He </PER> therefore, doubtful of the course of things, resolved <Action> to leave </Action> the country for <PLACE> America </PLACE>, where he hoped to meet with a refuge from the impending storm." [18]

     Explanation: In this context, there is no suggestion that the character is at the place repeatedly or forever. Hence "PUNCTUAL".

  3. "<PER> We </PER> <Action> embarked at </Action> <PLACE> Gravesend </PLACE> on the 9th of June, 1764, in the Anglicana, a ship carrying sixteen guns, and thirty-two men, burden about three hundred tons." [1]

Explanation: From the context available, there is no indication that this is a repeating interaction of the character with the place.

4. "<PER> We </PER> were detained there from the 17th to the 24th, when <PER> we </PER> recovered our anchor with some difficulty, and <Action> got clear of </Action> the <PLACE> harbour </PLACE>." [1]

   Explanation: The characters referenced "*We*" have been in the harbour for a few days, but it was likely for just one occasion.

5. "This was accordingly done, and the <PER> youth </PER> <Action> sailed to </Action> the <PLACE> East Indies </PLACE> as a writer in the company's service". [18].

   Explanation: In this context, there is no suggestion that the character is at the place repeatedly or forever. Hence "PUNCTUAL".

- **HABITUAL:** If the primary character and the place have an enduring, seasonal, or recurrent spatial relationship, it should be annotated as "HABITUAL".

  1. "<PER> John Stewart, Esq. </PER> was born in <PLACE> London </PLACE> in the year 1749." [18]

     Explanation: The relationship between the character and the place is labelled "HABITUAL" because the character's birthplace indicates a spatial relationship.

  2. "At the age of thirteen <PER> he </PER> was <Action> removed to </Action> the <PLACE> CharterHouse </PLACE>: even there his aversion to books continued; he neglected every opportunity for study, at times only composing moral essays or themes, which, as the master declared, elicited genius superior to that of any boy in the class." [18]

     Explanation: From the limited context available, the character seems to have a long-term relationship with the place.

  3. "The salary he enjoyed was extremely handsome, although it was not paid until a subsequent period, many years after <PER> he </PER> had <Action> quitted </Action> <PLACE> India </PLACE>, and as totally unexpected by him at the time the fortunate news arrived." [18]

     Explanation: In this context, it's suggested that the character was in the place for a longer period. Hence "HABITUAL".

  4. "He had fondly cherished an idea that his doctrines would have been received with enthusiasm by the English; and having long digested that theory, <PER> he </PER>

was prepared, when <Action> settled in </Action> <PLACE> London </PLACE>, to make it public." [18].

Explanation: The relation between the character and the place is labelled "HABITUAL" because it indicates a long-term relationship as the character settled in 'London'.

5. "On Ash Wednesday morn, this good and amiable man died, without a pang, at <PER> his </PER> residence in <PLACE> Northumberland-street </PLACE>." [18]

   Explanation: From the context available, there is an indication ["*residence*"] that this is a repeating interaction of the character with the place as it's the residence of the character. Hence "HABITUAL".

Moreover, for instances like "*She is the maid of the Mill*" [3] where a character is given a title based on a location, the appropriate annotation should be IN, HABITUAL. Furthermore, the temporal span annotation is unnecessary if the spatial relationship is "NO REL". As a general guideline, several months are often considered long-term, and this can be used as a heuristic when categorizing an instance as "HABITUAL". However, exercising your best judgment in assessing whether the interaction is genuinely long-term or short-lived is important, as context and specific details can influence this determination.

IsPartOfJourney

This binary annotation determines whether the referenced location, associated with the primary character (identified as [id = 0]), corresponds to a segment of that individual's journey. Specifically, if the *Spatial Relation* is categorized as IN, FROM, TO, NEAR, or THROUGH, the primary character's interaction with the place is considered part of their journey, and the annotation is assigned the value "TRUE". On the other hand, if the *Spatial Relation* is categorized as "NO REL" or the passage describes a journey that occurred in a prior period, distinct from the timeframe of the current narrative, then the mentioned place associated with the primary character is not considered part of the journey, and the annotation should be assigned the value "FALSE". This approach ensures a clear and consistent representation of spatial interactions, enabling precise tracking of the character's movement and contextualizing their journey within the narrative.

Furthermore, we also want to differentiate interactions between the characters and the place as part of the journey. On face value, this might appear as just marking any connection between the characters and the place of the sentence. Still, more precisely, this annotation asks the annotator to mark the travel relationship to the current narration travel. For example, consider the following sentence:

"<PER> Mr. Stewart </PER>, born in <PLACE> London </PLACE> in the year 1749; who made the tour of <PLACE> Scotland </PLACE>, <PLACE> Germany </PLACE>, <PLACE> Italy </PLACE>, and <PLACE> France </PLACE>, on foot, and ultimately settled in <PLACE> Paris

</PLACE>." [18]

In this case, the narration is describing <PER> Mr. Stewart </PER> journey. Within this narrative connection between the character and the place, <PER> Mr. Stewart </PER>, born in <PLACE> London </PLACE>" is not part of the journey. Still, the relationship is valid, so we would mark the relation as "FALSE" (concerning that IsPartOfJourney). However, the narration involves a tour that took place at some time previous to the current narrative time ("who made the tour of <PLACE> Scotland </PLACE>, <PLACE> Germany </PLACE>, <PLACE> Italy </PLACE>, and <PLACE> France </PLACE>, on foot, and ultimately settled in <PLACE> Paris </PLACE>"); we would hence mark this relationship as "TRUE".

Additionally, if the narrative meets any of the following conditions, the mentioned place is also excluded from the journey, and the annotation should be "FALSE": **Keywords and phrases indicating future intentions** play a crucial role in distinguishing planned or hypothetical travel from actual movement. Verbs such as *plan, intend, decide, prepare, arrange, schedule, aim, hope, and expect* explicitly signal intention, while modal verbs like *will, shall, would, should, and could* suggest actions that may take place in the future. Additionally, phrases such as *thinking about traveling to, going to, drawing on, obliged to, discussed visiting, contemplating a journey to, changed my plan, decided on continuing my journey by land, and set our sights on* provide further context for anticipated but unrealized travel. **Expressions of preparation** also contribute to understanding travel intent. Logistical arrangements, including *booking passages, packing belongings, securing travel documents, and arranging accommodations*, indicate concrete steps taken towards a journey. Financial considerations, such as *budgeting for the journey, exchanging currency, or financial planning related to travel expenses*, further reinforce the preparatory phase. Gathering information, such as consulting *maps, guides, or seeking advice about destinations and routes*, or referencing the *distance from the destination* (e.g., *fifty leagues from Constantinople*), also signals an anticipation of travel.

Time references offer another layer of indication for future travel. Words and phrases such as *next week, soon, upcoming, in the coming months, eventually, and afterward* act as direct indicators of a future timeline, while conditional statements, such as *If the weather permits, we will...* or *Should the opportunity arise, I plan to...*, suggest that travel is contingent upon specific conditions. **Dialogues or correspondence** also provide insight into travel intentions. Conversations where individuals discuss *where and when they intend to travel* or seek *permission to visit a place* indicate future movement. Similarly, *letters or messages detailing written plans or invitations related to future or past journeys* reinforce the notion of anticipated travel.

Emotional expressions associated with travel reveal further nuances of intention. Feelings of *excitement, anxiety, or anticipation regarding a planned journey* suggest an impending departure. Likewise, motivations such as the *desire to explore, the necessity of relocation, or obligations requiring travel* serve as underlying reasons for planned movement. Collectively, these linguistic markers help differentiate planned or hypothetical travel from actual journeys, offering critical insights for

historical itinerary reconstruction.

Within this category, there are two possible values: "TRUE" and "FALSE". Below illustrate with a few examples[1]:

- **FALSE:** The event sequence involving the character and the place has a connection but is not part of the journey. Please note that a passage may describe multiple events in a sequence, and the event containing the interaction between the place and the character might be one. The annotation should be "FALSE" as long as the event sequence is not part of the journey. To clarify this category, the illustrative instances are provided as follows:

  1. "<PER> Mr. Stewart </PER>, born in <PLACE> London </PLACE> in the year 1749." [18]

     Explanation: The relation between the character and the place is valid but not part of the journey.

  2. "<PER> He </PER> was removed to the <PLACE> CharterHouse </PLACE>: even there is a version to books continued; he neglected every opportunity for study, at times only composing moral essays or themes, which, as the master declared, elicited genius superior to that of any boy in the class." [18]

     Explanation: The relationship between the character and the place indicates a past journey, but it did not occur at the current narrative time.

  3. "His father was advised to send him abroad, in the hopes of removing his, indolent habits. This was accordingly done, and <PER> the youth </PER> <Action> sailed to </Action> the <PLACE> East Indies </PLACE> as a writer in the Company's service." [18]

     Explanation: The relationship of the character and the place is valid, but the narration involves a tour that took place at some time previous, and it did not take place at the current narrative time.

  4. "<PER>I</PER> was requested to visit an <PER>Armenian merchant</PER>, who had just <Action>arrived from</Action> <PLACE>Damascus</PLACE>." [21]

     Explanation: The relationship between the main character and the "*Armenian merchant*" is clearly valid in this context. However, there is no direct relationship between the main character and "*Damascus*", as the location is associated with another character's journey rather than the main traveler's itinerary.

  5. "<PER>I</PER> again changed <PER>my</PER> plan and decided on <Action>continuing</Action> <PER>my</PER> <Action>journey by</Action> land to <PLACE>Roustchouk</PLACE>." [20]

Explanation: The relationship between the character and the place is valid, but the narrative indicates a future journey ("*plan and decided*"), and there is no clear indication of reaching the place.

- **TRUE:** In a sequence of events, this event involving the character and the place connection is part of the journey which already happened at some point in the past (relative to the current narrative time). Examples include tours, travel, journeys of past events, etc. To clarify this category, the illustrative instances are provided as follows-

  1. "<PER> Mr. Stewart </PER> took <Action> leave of </Action> <PLACE> India </PLACE>, and <Action> traveled over </Action> near the whole of <PLACE> Persia </PLACE> and <PLACE> Turkey </PLACE> on foot." [18]

     Explanation: The sentence suggests that the character was in India sometime in the past; from there, he had traveled through the places ("*traveled over near the whole of*").

  2. "<PER> He </PER> made the <Action> tour of </Action> <PLACE> Scotland </PLACE>, <PLACE> Germany </PLACE>, <PLACE> Italy </PLACE>, and <PLACE> France </PLACE>, on foot, and ultimately <Action> settled in </Action> <PLACE> Paris </PLACE>." [18]

     Explanation: It's clearly stated in the text that the character referenced by "he" visited the places in the past ("*made the tour of*") and settled in Paris

  3. "<PER>I</PER> <Action>passed</Action> the morning of December 25th, <Action>opposite</Action> the village of <PLACE>Massoura</PLACE>." [39].

     Explanation: It is unclear in this excerpt whether the character is in *"Massoura"*. However, the character is inevitably in the opposite direction of *"Massoura"* (since he *passes the morning opposite of "Massoura,"* which indicates the character is near the place). Hence, this interaction is part of the journey.

  4. "<PER>He</PER> accordingly took a <Action>passage on</Action> board a ship <Action>bound to</Action> <PLACE>New York</PLACE>, and paid the Captain the sum required." [18]

     Explanation: The sentence describes the character embarking on a journey from one place to another. Additionally, the mention of the ship being bound for "*New York*" provides explicit confirmation that the character will reach the destination, justifying the travel connection between person and place entity.

  5. "<PER> He </PER> took his <Action> departure from </Action> <PLACE> Paris </PLACE> <Action> to </Action> <PLACE> England </PLACE>, unable to redeem his property in time through the baste he purposely made." [18]

     Explanation: The character referenced by "*he*" traveled from one place to another. This,

however, is not the current sequence in the narrative time. Hence this interaction is part of the journey ("*took his departure from*").

Furthermore, a useful way to think about the narrative tense is that if the passage refers to a travel that took place at some time previous to the current narrative time but part of the current travel in a sequence then the travel is "*part of the journey*". If `IsPartOfJourney=TRUE`, and the character is moving from `A` to `B` via `X/Y`, stays for some time or a few days in `Z`, and finally reaches the destination `B`, then follow these steps to construct the itinerary representing the character's actual movement correctly:

1. **Starting Point (FROM):** If the *Spatial Relation* is `FROM`, indicating the departure or start of the journey (such as `A`), then this place comes first in the itinerary.
2. **Intermediate Locations (THRU or NEAR):** If the *Spatial Relation* is `THRU` or `NEAR` (such as `X/Y`), then these places are placed in the middle of the itinerary.
3. **Stops or Stays (IN):** If the *Spatial Relation* is `IN`, indicating a stay at a place for a few days or a temporary stop (such as `Z`), then it should be included in the itinerary according to the sequence of the journey.
4. **Destination (TO):** If the *Spatial Relation* is `TO`, indicating the destination where the character has arrived (such as `B`), then it is placed at the end of the itinerary.

Therefore, the construction of the itinerary should be as follows: `A → X/Y → Z → B`.

However, if the spatial relation is labelled as "`NO REL`" then the "`IsPARTOfJOURNEY`" annotations need not be made.

Valid connection

This binary annotation indicates whether there is a relationship between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and other characters. The annotation should be "`FALSE`" if the characters have no relationship. Moreover, the first annotation serves as a crucial filtering step, aimed at curating a subset of examples characterized by high quality and devoid of entity tagging errors. However, the valid connection determines the existence of a relational link between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and other characters. If the primary characters within the monologue share a direct, indirect, or any form of connection with different characters, it warrants annotation as "`TRUE`". Conversely, in cases where no such relationship or connection exists among the characters, the suitable annotation for this field is "`FALSE`". For annotation purposes, we assign id=0 to the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and id=1 or 2 or 3 to other characters.

Within this category, there are two possible values: "TRUE" and "FALSE". Below are a few examples[1]:

- **FALSE:** The event sequence involves more than one character and might not have a direct, indirect or any form of connection between them. Please note that a passage may describe multiple events in a sequence of which the event containing the interaction between the characters might be one. The annotation should be "FALSE" if the event sequence does not indicate any relationship among them. To clarify this category, the illustrative instances are provided as follows-

  1. "And now, gentle reader, <PER id=0> I </PER> beg leave to withdraw for a season from public notice, bidding you farewell in one of <PER id=1> Statius' </PER> sweetest lines, in his seventh Bucolique: 'Tu benefice valerum "Non adequitus rere tu volumnis," Spoouactus dere, bumfolastri spoliatro'." [18]

     Explanation: In this context, it's clear that there is no valid relationship between the characters.

  2. " <PER id=1> Ibrahim d'el Arych </PER> turned back to the Oasis, which <PER id=0> we </PER> had left at some distance, in quest of an Arab who was better acquainted with this labyrinth of canals than our guides, who conceived the attempt to be too hazardous and difficult." [1]

     Explanation: In this context, no indication is that the characters referenced by "*we*" and "*Ibrahim d'el Arych*" have any direct or indirect connection. So, there is no relationship between them.

  3. "<PER id=0> Mr. Stewart </PER> observes, that authors, in all foreign countries, for want of thought, have no prudence in their doctrines or systems; they attack the guard-frame of the plant, instead of improving the soil, as exemplified by <PER id=1> Mira- beau </PER>, who overthrows all religious principles, without substituting any philosophic ones in their place; ." [18]

     Explanation: From the limited context available, the relation between the characters is not valid as no direct or indirect connection exists between the characters.

  4. "The Philosopher, however, was somewhat aroused from the calm and tranquil life <PER id=0> he </PER> passed in <PLACE> Cockspur-street </PLACE>, by the sudden arrival of Her late Majesty, <PER id=1> Queen Caroline </PER>, whose appearance in England, was the signal for riot and tumult among the lower classes of the population, who were glad of this pretext to Titter their seditious sentiments under covert of her cause." [18]

     Explanation: In this context, there is no clear indication that the characters referenced

by "*he*" and "*Queen Caroline*" have any direct or indirect connection. So, it's labelled as "FALSE".

- **TRUE:** In a sequence of events, this event involving the characters might have a direct, indirect or any form of connection. To clarify this category, the illustrative instances are provided as follows-

  1. "At the age of ten his <PER id=1> father </PER> sent <PER id=0> him </PER> to Harrow school." [18]

     Explanation: Both the characters ("*father*" and "*him*") are directly connected to each other.

  2. "<PER id=0>I</PER> was requested to visit an <PER id=1>Armenian merchant</PER>, who had just <Action>arrived from</Action> <PLACE>Damascus</PLACE>." [21]

     Explanation: In this context, the relationship between the main character and the "*Armenian merchant*" is clearly valid, as the main character intends to meet the other individual, establishing a direct interaction.

  3. "<PER id=0> I </PER> shall always reflect with great satisfaction on the adventures of a journey during which I enjoyed the pleasure of the society of my friend <PER id=1> Mr. Legh </PER>." [19]

     Explanation: Both the characters ("*I*" and "*Mr. Legh*") have a relationship (friendship) between them, which means they are directly connected to each other.

  4. "<PER id=0> I </PER> hope so impossible a summe to the greatest prince of Christendome to advance on the sudden, being I may name it to your <PER id=1> Tzarskoy Majesty </PER>, ten thousand poods of silver, to the value of above thirty hundred thousand rubles, was not demanded on purpose to have a pretext to deny the privileges, and by proposing an impossibility to refuse what is rational." [19]

     Explanation: Even though the characters do not have a direct relationship, they are connected in this monologue as the character referenced by "*I*" offers something to the character "*Tzarskoy Majesty*". So, they have a valid connection.

Moreover, A useful way to think about the valid connection is that if the character references family members, friends, travel partners, colleagues, or shares a monologue with the primary character (knows the character personally), the connection is "TRUE". Suppose the primary character refers to another character as a reference or shares that character's dialogue but personally has no relationship with the character. In that case, the annotation should be "FALSE". Furthermore, if the spatial relation is labelled as "NO REL", the "*Valid Connection*" annotations need not be made.

Travel connection

This categorical annotation indicates the travel relationship between the characters. There are four categories as mentioned in Table 5.1 and definitions are given in Table 5.4.

Table 5.4: Short description of travel connection's subtask

Category	Short description
IN GROUP	Characters traveling together from the beginning.
JOIN GROUP	The character joins the traveler's group/ another character from the middle of traveling.
EXIT GROUP	The character leaves the traveler's group/ another character in the middle of traveling.
NO REL	There is no relationship between the characters in terms of travel.

Additionally, this annotation filters the type of relationship (e.g. *Travel relationship*) between the primary character and other characters. Furthermore, the following annotations should only be applied when the valid connection has been designated as "TRUE". Various values can be assigned to this annotation, to clarify these categories, the illustrative instances are provided as follows[1]:

- **IN GROUP:** The category indicates whether a character starts a journey with a group or another character defining the travel connection. This association can be explicitly indicated by mentions of "*Companion*," "*Guide*," or "*Fellow Traveler*", shared actions, destinations, and experiences, or it may need to be deduced. The key determination to be made is whether the character's position falls within the beginning of the journey. While lexical indicators can aid in this determination when available, this relationship may need to be inferred from the context when such cues are absent. To clarify these connections, the illustrative instances are provided as follows-

  1. "<PER id=0>I</PER> shall always reflect with great satisfaction on the adventures of a journey during which I enjoyed the pleasure of the society of my friend <PER id=1>Mr. Legh</PER>" [19].

     Explanation: It's directly indicated in the text that the character referred to by "*I*" enjoys the journey with another character referred to as "*Mr. Legh*", meaning they travel together.

  2. "<PER id=0>I</PER> was provided with a ferman from the Porte, and accompanied by a <PER id=1>Janissary, Mustafa</PER>, belonging to the English embassy." [19]

     Explanation: In this context, the main character's indication of being "*accompanied by*" suggests that they are traveling together, establishing a shared journey.

  3. "<PER id=0>Our</PER> <PER id=1>Arab guides</PER> had endeavoured to

alarm <PER id=0>us</PER> as to the consequences of bathing in these pestiferous waters." [19]

Explanation: The limited context suggests they are traveling together, as the main traveler's group refers to the other individual as their guide, implying a shared journey.

4. "<PER id=0>we<</PER> arrived about ten at night; <PER id=0>our</PER> <PER id=1>drogoman, Abou Doaud</PER>, had worked up <PER id=1>his</PER> imagination that <PER id=1>he</PER> calculated on indulging in a repose the next morning, but it turned out otherwise." [39]

Explanation: The limited context suggests they are traveling together, as the primary traveler's group refers to the other individual as their drogoman, implying a shared journey.

- **JOIN GROUP:** This category indicates whether a character joins a group or another character in the middle of the journey and continues traveling with them. "JOIN GROUP" typically conveys an "IN GROUP" relationship as well, but offers a more detailed perspective on the characteristics of that relationship. Importantly, this category stipulates that the character's starting point is distinct from the group or another character's starting point, which they are identified as traversing. Furthermore, this applies when the group or another character starts at place X (which is not Y), traverses through place Y from where a character joins, and concludes at place Z (which is not Y). The character should not have to start the journey at place X, but should have to join the group or another character while they traverse through place Y. As this category implicitly suggests that the character is in the group, the pivotal differentiation for "JOIN GROUP" is that the character is presumed to be in the group or with another character from the middle of the journey, whereas "IN GROUP" indicates starting the journey together. To clarify these connections, the illustrative instances are provided as follows:

  1. "After dark, <PER id=1>Mr. Ritchie</PER>, accompanied by <PER id=0>our</PER> friends, <PER id=2>Dr. Dickson</PER>, and <PER id=3>Messrs. Carstensen</PER>, arrived." [41]

     Explanation: The character referred to as "*Mr. Ritchie*", "*Dr. Dickson*" and "*Messrs. Carstensen*" joins the main traveler's group after dark, which means in the middle of their journey. So the label is "JOIN GROUP".

  2. "At the port of Larnica, in Cyprus, where, at the earnest request of a <PER id=1>Greek priest</PER>, who came to pay <PER id=0>me</PER> a visit, fifty pilgrims bound to Jerusalem, were taken on board." [19]

     Explanation: The limited context indicates that the character referred to as "*Greek priest*" joins the main travelers group and continues their journey. Hence, "JOIN GROUP".

3. "In a few days, our party was increased by the arrival of the Honble. <PER id=1>Captain Irby</PER> and <PER id=2>Captain Mangles</PER>, both of the Royal Navy, who had been absent for a short time to visit Bethlehem." [19]

   Explanation: Similar to the above example, the character referred to as "*Captain Irby*" and "*Captain Mangles*" joins the group and they continue traveling together.

4. "As <PER id=0>we</PER> made <PER id=0>our</PER> way through the dense forest, <PER id=0>we</PER> encountered a man named <PER id=1>Marcus</PER> traveling alone. After a brief conversation, it became clear that <PER id=0>we</PER> were heading in the same direction. From that point on, <PER id=1>Marcus</PER> walked alongside <PER id=0>us</PER>, sharing <PER id=1>his</PER> knowledge of the land as <PER id=0>we</PER> continued toward the distant mountains."

   Explanation: In this context, the phrases "*walked alongside us*" and "*from that point on*" imply that "*Marcus*" has become part of the group in the middle of their journey. Hence, "JOIN GROUP".

- **EXIT GROUP:** This category indicates whether a character leaves a group or another character in the middle of their journey. "Exit GROUP" typically can convey an "IN GROUP" or "JOIN GROUP" relationship as well, but indicate a different destination. Importantly, this category stipulates that the character's ending point is distinct from the group or another character's endpoint, and they are identified as the destination. Furthermore, this applies when the group or another character starts at place X (which is not Y), traverses through place Y where a character may take leave, and concludes at place Z (which is not Y). The character may start the journey at place X or join the group or another character while they traverse through place Y, but take leave before they reach place Z. As this category implicitly suggests that the character is in or joins the group, the pivotal differentiation for "EXIT GROUP" is that the character is presumed to be left the group or another character in the middle of the journey, whereas "IN GROUP" or "JOIN GROUP" indicates end the journey together. To clarify these connections, the illustrative instances are provided as follows:

  1. "<PER id=0>Our</PER> servant, <PER id=1>Nicolo</PER>, was sent with all <PER id=0>our</PER> baggage and valuables, to await <PER id=0>our</PER> return at Acre." [19]

     Explanation: It's directly indicated in the text that the character referred to by "*Nicolo*" left the travelers' group.

  2. "<PER id=1>Ibrahim d'el Arych</PER> turned back to the Oasis, which <PER id=0>we</PER> had left at some distance," [39]

     Explanation: In this context, the character "*Ibrahim d'el Arych*" leaves the group to follow

a different path, signalling his departure from the primary journey. So, it's labelled as "EXIT GROUP".

3. "Part of <PER id=0>our</PER> caravan, viz. the <PER id=1>General</PER> and <PER id=1>his</PER> attendants, the first Dragoman, and the Sheik, or chief of <PER id=0>our</PER> Bedouins, left <PER id=0>us</PER> here, and returned to Cairo." [42]

   Explanation: In this context, it's clearly indicated that the character referred to by "*General*" and his companions leaves the group to follow a different path.

4. "On the 22d March, 1819, after repeated delays, every thing being at length in readiness for <PER id=0>our</PER> departure, the camels, twenty-two in number, arrived, and carried <PER id=0>our</PER> goods to the Desert beyond the Meshea, where <PER id=1>Mr. Ritchie</PER> preferring to remain in the town." [41]

   Explanation: The limited context indicates that "*Mr. Ritchie*" leaves the group and prefers to stay in town. Hence,"EXIT GROUP".

- **NO REL:** Identifying character pairs that occur together in the text without explicitly implying a travel connection between them is one of the complicated tasks among other travel connections. The term "NO REL" indicates the absence of such a connection, compared to scenarios with insufficient information to determine the link between the characters (C). The concept of "NO REL" can be applied when a character is mentioned in the journey. Yet, it remains uncertain whether the character has been traveling with the group or the primary character, or if epistemic narrative ambiguity exists in the text, where the narrator or characters are unsure of the connection between the primary character and other characters. As demonstrated in the examples above, the connections between the primary character and other characters are frequently subtly implied, heavily relying on the reader's inference rather than explicit spatial cues within the text.

  Furthermore, when none of the previously described travel connections apply, the appropriate label is "NO REL". The "NO REL" label applies in various scenarios, including but not limited to - (i) When the text unambiguously indicates no discernible connection between the characters. (ii) When there is insufficient information available in the text to make a judgment about the travel connections between the characters. (iii) When the travel connections implied in the text are speculative or hypothetical. (iv) When the character is in the journey-related narrative, but it remains unclear if they ever travel together. (v) When it is definitively established that the character is not within the group or with the primary character. (vi) When the text explicitly conveys uncertainty regarding the character's connection with the travel group or the primary character. To clarify this category, the illustrative instances are provided as follows:

1. "At the age of ten his <PER id=1> father </PER> sent <PER id=0> him </PER> to Harrow school." [18]

   Explanation: From the context, it's clear that both the characters have a relationship but do not have any travel connection. So, it's labelled as "NO REL".

2. "To effect the primary object, <PER id=0> Mr. Stewart </PER> embraced an opportunity of quitting the Company's service, and entering into that of <PER id=1> Hyder Ally </PER>, in whose employment <PER id=0> he </PER> remained for some time, acquiring the good opinion and respect of this despot, who had sense to estimate the full worth of <PER id=0> Mr. Stewart's </PER> abilities." [18]

   Explanation: In this context, it's clear that both the characters have a relationship but no indication of their travel connection. Hence, "NO REL".

3. "The commander was <PER id=1> Capt. John Stewart </PER>; the price of <PER id=0> our </PER> passage to Turkey sixty guineas." [1]

   Explanation: From the limited context available, it's not clear whether the character referred to as "*Capt. John Stewart*" is traveling together with the group. So, it's labelled as "NO REL".

4. "<PER id=0> We </PER> had been advised to carry with <PER id=0> us </PER> money, for our' journey, in crown-pieces of silver, called imperial tallerie, from Leg <PER id=1> Mr. Rutherfurd </PER>, an English merchant, accepted <PER id=0> our </PER> bills on a banker in London." [1]

   Explanation: In this context, there is no indication that the character referred to as "*Mr. Rutherfurd*" is traveling with the group. Hence, "NO REL".

Furthermore, if the characters are not traveling or have no sign of any journey in the monologue, but the valid connection is "TRUE", then the annotation should be "NO REL". If the valid connection is labelled as "FALSE", then the "*Travel Connection*" annotations need not be made.

Chapter 6

Data Preparation

This chapter offers a comprehensive discussion of the data utilized for this project, encompassing the data pre-processing steps, corpus preparation, and manual annotation procedures that were prepared during an 18 credit research-oriented project [36]. Each aspect is examined in detail to provide a clear understanding of the methodologies employed to ensure data quality and suitability for the research objectives.

6.1 Data and their Digital Format

Travel accounts from the 19th century pose distinctive challenges and opportunities for researchers. Despite offering detailed descriptions of landscapes, cities, and societies, as illustrated in example-1 [Chapter 1], accurately extracting travel routes can prove to be a challenging task. This complexity is amplified by frequent references in the texts to locations that are not included in the itineraries, such as the mention of "Spanish towns" and "St. Philip's castle" in example- 1, as these places were not physically visited by the travelers and the information is presented without a structured format. To assess the travel connection between characters and locations, this project utilizes various textual references, consisting of around 120 pages: the compilation of historical travelogues of the Ottoman Empire (In English language).

Numerous 19th-century historical travelogues have been digitized and are accessible in various research-friendly formats (Archives, EPUB, PDF Scans etc.) through sources like Project Gutenberg, HathiTrust Digital Library, Internet Archive, Google Books, and the British Library Online Collections. For this project, 15 randomly selected digital versions (PDF scans) of historical travelogues of the Ottoman Empire (In English language) were gathered from Google Books, which provides widely available public domain texts for public use [Table 6.1] [9]. These are digital reproductions of the original printed books, capturing every detail as it appears on the physical pages. [1]

[1]Note: The data used for corpus preparation are in the public domain, eliminating potential copyright infringements.

Table 6.1: Data Source List

No.	**Title (Year of Publication)**	**Author**
1	Travels In Asia Minor , And Greece: Or An Account Of A Tour Made At The Expense Of The Society Of Dilettanti (1817) [1].	Richard Chandler
2	Travels in Various Countries of Europe, Asia and Africa (1817) [43].	E. D. Clarke
3	KARAMANIA: A Brief Description of the South Coast of Asia Minor and the Remains of Antiquity (1817) [38].	Francis Beaufort
4	Description Of The View Of Athens, And Surrounding Country (1818) [40].	Henry Aston Barker
5	Travels in Egypt, Nubia, Holy Land, Mount Libanon, and Cyprus, in the Year 1814 (1818) [44].	Henry Light
6	Journey from Moscow to Constantinople, in the years 1817, 1818 (1819) [19].	William Macmichael
7	Memoir on the Commerce and Navigation of the Black Sea, and the Trade and Maritime Geography of Turkey and Egypt (1819) [45].	Henry A. S. Dearborn
8	Travels in Egypt, Being a Continuation of the Travels in the Holy Land in 1817-18 (1819) [39].	Count de Forbin
9	The Military Costume of Turkey. Illustrated by a Series of Engravings from Drawings Made on the Spot: Dedicated by Permission to His Excellency the Minister of the Ottoman Porte to His Britannic Majesty (1819) [46].	John Heaviside Clark
10	A Voyage up the Persian Gulf, and a Journey Overland from India to England, in 1817 (1819) [47].	Lieutenant William Heude
11	A Narrative of a Journey into Persia and Residence at Teheran, Containing a Descriptive Itinerary from Constantinople to the Persian Capital (1820) [20]	J. M. Tancoigne
12	A Narrative of Travels in Northern Africa, in the Years 1818, 19, and 20 (1821) [41].	Captain G. F. Lyon
13	The Life and Adventures of the Celebrated Walking Stewart: Including His Travels in the East Indies, Turkey, Germany, & America. By a Relative. With a Portrait (1822) [18].	William Thomas Brande
14	Travels Along the Mediterranean, and Parts Adjacent; In Company with the Earl of Belmore, During the Years 1816-17-18 (1822) [21].	M.D. Robert Richardson
15	Travels in the Countries Between Alexandria and Paraetonium, the Lybian Desert, Siwa, Egypt, Palestine, and Syria, in 1821 (1822) [42].	Johann Martin Augustin Scholz

6.2 Data Extraction Procedure

The 'pdfplumber' [Section 4.4] library is used to automate the extraction of text from the historical travelogues of the Ottoman Empire's (in English language) PDF Scan files and saves the extracted content into a Word document with the help of the 'python-docx' library. The primary objective is to convert a scanned 19th-century travelogue into an editable format while addressing common Optical Character Recognition (OCR) issues inherent in historical documents. In algorithm 1, the algorithm is provided for the data extraction process.

Initially, a new Word document 'doc' is created. The algorithm opens the PDF file within a context manager to ensure proper resource management and iterates through each page starting from the second one (skipping the first page, which might contain non-essential information like a cover or table of contents). For each page, it extracts individual words using '`page.extract_words()`' with parameters fine-tuned for OCR correction—such as '`expand_ligatures=True`' (Historical texts often use ligatures or combined letters, which can cause errors in text extraction) to handle historical ligatures by setting '`expand_ligatures=True`' helps in converting these to standard characters and '`use_text_flow=True`' ensures that the text is extracted in the order it would be read, which is important for documents with multiple columns or non-standard layouts. Furthermore, Adjusting '`x_tolerance`' and '`y_tolerance`' helps in grouping characters and words accurately, which is crucial when dealing with scanned documents that might have irregular spacing. Although the extracted words ('`word_tokens`') are not used subsequently in the code, they serve future text processing needs in the section 6.4 for NER model. Simultaneously, the full text of each page is extracted using '`page.extract_text()`' with similar parameters to ensure consistency in text handling. The extracted text is then added to the Word document, prefaced with a page number for reference using '`doc.add_paragraph(f"Page i")`' and '`doc.add_paragraph(text)`'. After processing all relevant pages, the document is saved via '`doc.save("TravelAccount.docx")`'. This comprehensive process effectively transforms the scanned PDF into an editable and searchable Word document, facilitating further analysis and research on the historical text while mitigating issues related to OCR errors and the complexities of 19th-century print layouts.

Algorithm 1 Convert PDF to Word Document

```
Create a new Word document
doc ← Document()
Open the PDF file "TravelAccount.pdf" using pdfplumber
with pdfplumber.open("TravelAccount.pdf") as pdf:
Loop through each page in the PDF
for each page in pdf.pages with index i do
    Skip the first page (start processing from the second page)
    if i ≥ 1 then
        Extract words from the page using OCR correction with specified parameters
        word_tokens ← page.extract_words(
                x_tolerance = 1,
                y_tolerance = 2,
                keep_blank_chars = True,
                use_text_flow = True,
                horizontal_ltr = True,
                vertical_ttb = True,
                extra_attrs = [],
                split_at_punctuation = False,
                expand_ligatures = True)
        Extract plain text from the page using OCR correction with specified parameters
        text ← page.extract_text(
                x_tolerance = 1,
                y_tolerance = 2,
                layout = False,
                use_text_flow = True,
                keep_blank_chars = True,
                split_at_punctuation = False,
                expand_ligatures = True,
                horizontal_ltr = True,
                vertical_ttb = True)
        Add a paragraph to the Word document indicating the page number
        doc.add_paragraph("Page i")
        Add the extracted text to the Word document as a new paragraph
        doc.add_paragraph(text)
    end if
end for
Save the Word document as "TravelAccount.docx"
```

6.3 Pre-processing Data

Despite the remarkable performance of *pdfplumber*, the extracted text has some issues. One of them is the inadvertent splitting of lines in a text, a common issue when processing OCR outputs from scanned documents like 19th-century travelogues. This issue has been addressed by a function [Algorithm 2] which is designed to correct words that have been inadvertently split across lines in a text. It starts by downloading an English vocabulary list from a GitHub repository

and reading it into a Python set called '`english_vocabulary`' for efficient word lookup. The function then checks whether a given word exists in this vocabulary set. The main function, '`fix_word_continuation(text)`', uses a regular expression to identify sequences of word characters that are split by newline characters—indicative of words broken across lines. It processes these matches in reverse order to avoid altering the indices of yet-to-be-processed text portions, which could happen if the text length changes. For each matched sequence, it reconstructs the probable complete word by removing all whitespace (including newlines) within that segment and then checks if this reconstructed word exists in the English vocabulary (case-insensitive). If it does, the code replaces the original split word in the text with the corrected word and adds a new line character after it to maintain the text's formatting. This systematic approach effectively repairs split words caused by line breaks in the OCR-processed text, enhancing the accuracy and readability of the document for further analysis or processing. In algorithm 2, the algorithm is provided which is used to correct words that have been split across lines in OCR-processed text.

Algorithm 2 Fix Word Continuation in Text

Download the English vocabulary file from the specified URL
`https://raw.githubusercontent.com/dwyl/english-words/master/words_alpha.txt`
Open the downloaded file and load the English vocabulary into a set
with open("/content/words_alpha.txt") as f:
$english\_vocabulary \leftarrow$ set($f.read().split('\n')$)
function IS_WORD_IN_ENGLISH_VOCABULARY(word)
    return $word \in english\_vocabulary$
end function
function FIX_WORD_CONTINUATION(text)
    Find all matches of words split across lines using regex
    $results \leftarrow [(m.start(0), m.end(0))$ **for** m **in** $re.finditer(r'\w + (\n\w+)+', text)]$
    Process matches in reverse order
    for each res in $results[:: -1]$ **do**
        Combine the split parts of the word into a single string
        $probable\_word \leftarrow$ "".$join(text[res[0] : res[1]].split())$
        Check if the combined word exists in the English vocabulary
        if is_word_in_english_vocabulary($probable\_word.lower()$) **then**
            Replace the split word with the combined word in the text
            $text \leftarrow text[: res[0]] + probable\_word + ' \n' + text[res[1] :]$
        end if
    end for
    return $text$
end function

6.4 Corpus Preparation

The corpus for this study comprises 120 pages (~21,674 words) extracted from 15 historical travelogues of the Ottoman Empire (In English language) [1] [48]. These texts provide firsthand

accounts of travellers' experiences and observations, offering rich material for analyzing person and place entities.

To identify and extract sentences containing person and place entities, *spaCy's* [Section 4.3.2] NER model was utilized in conjunction with the *SpanMarker* [Section 4.3.1] model. Chapters containing the highest number of person and place entities were selected for the corpus to ensure that the analysis includes abundant person-place relationships within the full context of the narrative, thereby aiding in visualizing the itinerary [Table 6.2]. In algorithm 3 and 4, the algorithm is provided which is used for entity extraction.

Algorithm 3 Find Location Sentences in Text (Part 1)

```
Import necessary libraries:
from transformers import pipeline
from spacy.lang.en import English
import spacy
import pytextrank
from spacy.tokens import Span
from span_marker import SpanMarkerModel
Load the Hugging Face NER model
model ← SpanMarkerModel.from_pretrained(
        "tomaarsen/span − marker − roberta − large − fewnerd − fine − super",
        model_max_length = 1024,
        marker_max_length = 128,
        entity_max_length = 16)
Initialize the SpaCy tokenizer
tokenizer ← English().tokenizer
Define a custom pronoun scrubber function
@spacy.registry.misc("pronoun_scrubber")
function PRONOUN_SCRUBBER
    function SCRUBBER_FUNC(span: Span)
        for each token in span do
            if token.pos_ not in ["PRON"] then
                break
            end if
            span ← span[1 :]
        end for
        return (span.start_char, span.end_char)
    end function
    return scrubber_func
end function
Load the SpaCy generic model
spacy_model ← spacy.load("en_core_web_sm")
Add the TextRank pipeline with the custom scrubber
spacy_model.add_pipe(
        "textrank",
        config = {"scrubber" : {"@misc" : "pronoun_scrubber"}})
```

Algorithm 4 Find Location Sentences in Text (Part 2)

```
function FIND_LOCATION_SENTENCES(text)
    Process the input text using the SpaCy model
    doc ← spacy_model(text)
    Extract sentences from the processed document
    sentences ← list(doc.sents)
    Create a dictionary to map spans to sentence indices
    span_to_sentence_dict ← {}
    for each sent in sentences with index i do
        span_to_sentence_dict[(sent.start_char, sent.end_char)] ← i
    end for
    function FIND_SENTENCE_IDX(start_char)
        for each k in span_to_sentence_dict do
            if k[0] ≤ start_char < k[1] then
                return span_to_sentence_dict[k]
            end if
        end for
        return −1
    end function
    Initialize a dictionary to store entities
    entities ← {}
    Extract named entities using SpaCy's NER
    for each word in doc.ents do
        span ← (word.start_char, word.end_char)
        if word.label_ == "GPE" or word.label_ == "PER" then
            sentence_idx ← find_sentence_idx(span[0])
            if sentence_idx ≠ −1 then
                entities[span] ← sentence_idx
            end if
        end if
    end for
    Extract named entities using Hugging Face's NER model
    length ← 0
    for each sentence in sentences with index sentence_idx do
        span_marker_entities ← model.predict(sentence.text)
        for each entity in span_marker_entities do
            span ← (
                    entity['char_start_index'] + length,
                    entity['char_end_index'] + length)
            if "location" in entity['label'] or "person" in entity['label'] then
                entities[span] ← sentence_idx
            end if
        end for
        Update the length counter
        length ← length + (sentence.end_char − sentence.start_char + 1)
    end for
    return entities, sentences
end function
```

The provided algorithm 3 is designed to identify and extract sentences containing person and location entities from a given text by combining *spaCy's* NER capabilities with the *SpanMarker* model from Hugging Face. It initializes the *SpanMarker* model for enhanced entity detection and sets up a *spaCy* pipeline with a custom "pronoun scrubber" to refine text spans. The code processes the input text, tokenizes it into sentences, and uses *spaCy* to detect entities labeled as "GPE" (Geo-Political Entity) or "PER" (Person). It then employs the *SpanMarke*r model to predict entities in each sentence, adjusting character indices to align with the original text. The result is a collection of entities mapped to their corresponding sentences, facilitating the analysis of person-place relationships within the narratives, which is particularly useful for visualizing itineraries in historical travel literature.

Table 6.2: Travelogue Corpus

BookID	Author	Year of Publication	Chapter	Pages	Words
001	William Thomas Brande [18]	1822	-	3-12	~3834
002	Henry Aston Barker [40]	1818	-	3-12	~2433
003	Richard Chandler [1]	1817	1-2	1-8	~2006
004	William Macmichael [19]	1819	4	187-207	~4845
005	Francis Beaufort [38]	1817	10	194-211	~2651
006	E. D. Clarke [43]	1817	4	199-213	~3577
007	Henry A. S. Dearborn [45]	1819	-	272-294	~4694
008	Count de Forbin [39]	1819	-	1-14	~6311
009	John Heaviside Clark [46]	1819	-	25-41	~1754
010	Johann Martin Augustin Scholz [42]	1822	-	1-11	~5429
011	Lieutenant William Heude [47]	1819	4	52-69	~6532
012	Henry Light [44]	1818	4	101-118	~4446
013	Captain G. F. Lyon [41]	1821	2	38-75	~6274
014	M.D. Robert Richardson [21]	1822	19	238-258	~5371
015	J. M. Tancoigne [20]	1820	34	386-394	~2017

6.5 Post-processing Data

Each document is transformed into an XML-like format for the annotation task, including metadata and the main content. The structure is defined by specific tags:

1. `<BookID>`: A unique identifier for each book.
2. `<Page>`: The original page number from the source text.
3. `<Author>`: The author's full name.
4. `<Title>`: The complete title of the work, including any subtitles or additional descriptors.
5. `<Volume>`: If the book has any volume number.

6. <Edition>: If the book has any edition number.
7. <Published>: The year of publication.
8. <Publication>: The publisher information.
9. <Other_Travellers>: If other travellers are accompanied by the main traveller and the names mentioned in the cover page, preface or abstract.
10. <Chapter_Title>: The title of the chapter with chapter number (if present).
11. <Chapter_Summary>: Any subtitles or additional descriptors of the chapter (if present).
12. <Data>: The extracted textual content from the page.

This structured approach allows efficient data management and retrieval, allowing targeted analyses based on author, publication date, or other metadata attributes. Including page numbers and book IDs facilitates traceability back to the original sources, ensuring transparency and reproducibility in the research.

6.6 Manual Annotation

The dataset for this study consists of excerpts sourced from English travelogues documenting journeys through the Ottoman Empire. The annotation process was systematically designed based on the guidelines developed in this research, ensuring consistency and accuracy in entity recognition and classification [Chapter 5].

To automate entity tagging, the GPT-4 model was employed to identify **Person (PER)** and **Place (PLACE)** entities within the text. Zeghidi et al. explore the evaluation of Named Entity Recognition (NER) using few-shot prompting techniques with Large Language Models (LLM) such as GPT-3.5, GPT-4 and GPT-4o [49]. Their study investigates the effectiveness of leveraging minimal annotated examples to guide the model's performance in entity recognition tasks. By employing few-shot prompting, the authors demonstrate how large language models can achieve competitive results on NER benchmarks, even with limited training data. This research incorporates the few-shot prompting concept while leveraging GPT-4 for the NER task. The automated approach eliminated the need for extensive manual tagging while significantly enhancing annotation efficiency. However, in cases where the model failed to recognize a place or person entity correctly, the annotator manually verified and assigned the appropriate label.

The entity tagging process involved applying GPT-4 to detect all person and place mentions, embedding inline <PER> and <PLACE> tags into the original text. The sample dataset comprised passages containing at least one character and one location, covering a total of 120 pages of annotated data. Following the initial tagging, the annotator conducted a meticulous verification and correction process, ensuring that all entities were accurately identified and appropriately

classified. Additionally, the annotator assigned classes to the identified entities, specifying relevant attributes such as **ValidConnection**, **TravelConnection**, **TemporalSpan**, and **IsPartOfJourney** to reflect their roles in the travel narrative.

Beyond entity recognition, the annotator also identified *Action keywords*, which denote movements or spatial relationships (e.g., "departed from," "arrived at"). Each `<Action>` tag was assigned a corresponding **SpatialRelation** class, ensuring that the traveler's movements were systematically captured and structured within the dataset.

The annotation process commenced with an initial pilot annotation phase, during which preliminary annotations were conducted to assess the guidelines' feasibility. After the first round of annotations, an annotation document was created and iteratively refined in subsequent pilot annotation rounds. This document outlined the annotation tasks, defined the classification categories for each entity and action type, and included intuitive real-world examples for each category to enhance clarity and standardization. This phase leveraged the MAMA sub-cycle concept to iteratively refine the annotation framework and ensure alignment with the study's objectives [Section 4.1.1].

As the annotation process progressed, the guidelines were continuously refined based on emerging challenges and annotation inconsistencies. The finalized annotation guideline document served as the standard reference, ensuring uniform application of entity and action classifications across the dataset. This manually annotated dataset serves as a gold annotation, providing a high-quality benchmark for evaluating the accuracy and effectiveness of automated entity tagging and relationship extraction systems.

To conclude, this chapter has provided a comprehensive discussion of the data used for this project, detailing the data preprocessing steps, the corpus preparation, and the manual annotation process. By meticulously extracting and refining text from 15 digitized historical travelogues of the Ottoman Empire (In English language) using tools like *PDFPlumber*, and employing *spaCy's* Named Entity Recognition model in conjunction with the *SpanMarker* model to identify sentences rich in person and place entities, the corpus has been effectively prepared. Furthermore, the manually annotated dataset functions as a gold-standard reference, offering a high-quality benchmark to assess the precision and efficacy of automated systems for entity tagging and relationship extraction. This structured and thorough approach ensures that the dataset is both reliable and suitable for analyzing person-place relationships within the narratives, thereby facilitating the visualization of itineraries and enhancing the understanding of narrative dynamics in historical travel literature.

Chapter 7

Implementation

This chapter aims to provide a detailed overview of the practical steps involved in implementing annotation guidelines through an automated workflow, utilizing advanced LLMs as the main tools. A well-structured methodology has been developed to incorporate data preparation, model selection, and evaluation. This guarantees that the final system can effectively annotate historical travelogues in accordance with the guidelines. This process not only fulfills the fundamental objectives outlined in previous chapters but also demonstrates the adaptability of modern LLMs in handling specialized tasks such as identifying valid itineraries and distinguishing visited locations from casual references.

The implementation intends to:

1. **Automate the annotation process through domain specialized prompt crafting:** Specializing Large Language Models (LLM) involves tailoring general-purpose LLMs using specific domain contextual information, enhanced by domain-specific knowledge, aligned with the domain's goals, and guided by domain-specific limitations [50, p 1]. Furthermore, prompts or task-specific input texts created to guide content generation by the LLM help set expectations for the desired model responses. There are two main approaches: *(a) Discrete Prompt* involves providing natural language instructions specific to the task to prompt LLMs for domain-specific knowledge, and *(b) Continuous Prompt* utilizes learnable vectors to prompt LLMs without the need for manually crafted text instructions [Figure 7.1] [50, p 6, 12]. This study utilizes *Discrete Prompt* by employing two-shot and few-shot prompting techniques to systematically label person–place co-occurrences in historical travel texts, identifying whether such locations are part of an actual itinerary [Section 4.5].

2. **Maintain scalability Without additional training:** Large Language Models (LLM)s demonstrate exceptional proficiency in leveraging a limited number of input-output examples ("shots") provided in context during inference, enabling them to perform complex tasks without requiring weight updates. This approach ensures broad adaptability across di-

verse datasets and textual domains by exclusively utilizing pre-trained language models in conjunction with meticulously crafted prompts.

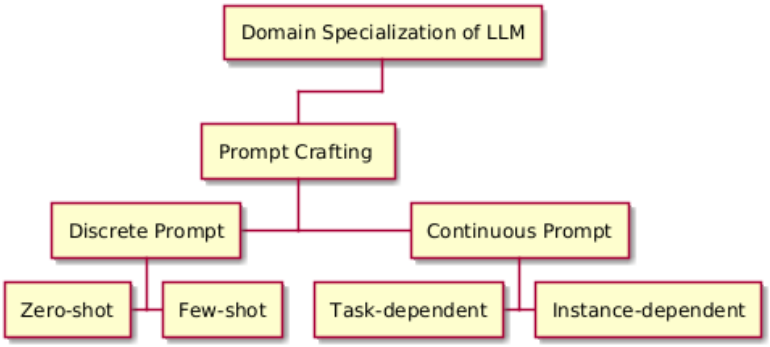


Figure 7.1: Domain Specialization of LLM

The method sidesteps the need for extensive annotated training data or specialized retraining cycles by emphasizing prompt-crafting techniques over fine-tuning. Instead, it directly exploits the general language understanding capabilities of the chosen LLMs while still adhering to the annotation guidelines in earlier chapters [Chapter 5]. The annotation guidelines, detailed in previous chapters, focus on distinguishing whether places mentioned in historical travelogues are genuinely part of a traveler's journey. The primary categories include:

(i) IsPartOfJourney: Confirms that a person–place co-occurrence is integral to the travelogue itinerary based on context and spatial indicators.

(ii) Valid Connection: Indicates a relationship between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and other characters or not.

(iii) Travel Connection: Conveys the travel relationship between the characters.

The guidelines also describe criteria for identifying irrelevant or generic mentions, ensuring that only meaningful place references are tagged as part of the traveler's route.

Three state-of-the-art LLM models GPT by OpenAI, LLaMA, and Mistral through the Ollama platform were chosen for their diverse strengths, even when employed without additional training or fine-tuning:

OpenAI GPT (Generative Pre-trained Transformer) Models: GPT-based LLM models, developed by OpenAI, are renowned for their advanced natural language processing capabilities. [29]. It is available in various versions optimized for instruction-following tasks and text completion. While OpenAI GPT is not open-source, its API-based access allows researchers to leverage its

powerful generative and contextual understanding abilities. This research uses the OpenAI GPT model through its API, enabling efficient deployment and experimentation. The model's ability to generate coherent and contextually relevant text makes it well-suited for entity tagging, relationship extraction, and text analysis on historical travelogue datasets. Using OpenAI GPT ensures high-quality performance in processing and analyzing complex linguistic data. Various versions of the OpenAI GPT model, including GPT-4o,GPT-4 Turbo, and GPT-3.5 Turbo, were utilized in the experiment.

LLaMA (Ollama version): The Meta Llama 3 multilingual LLM is a pre-trained and instruction-tuned generative model that is open-source [30]. In this research, the quantized version of the Llama model is utilized through the *Ollama platform*, enabling efficient deployment and experimentation. This approach leverages the model's multilingual capabilities and generative strengths to address the specific requirements of the study, particularly in processing and analyzing the historical travelogue datasets. Its architecture supports local inference while maintaining competitive accuracy, making it a viable option for smaller-scale implementations.

Mistral (Ollama version): Mistral 7B (LLM) is distributed under the Apache license, making it an open-source model [33]. It is available in two variants: an instruct version, designed for instruction-following tasks, and a text completion version, tailored for generating coherent and contextually relevant text. This research employs the quantized version of the Mistral 7B model using the Ollama platform. This approach ensures efficient deployment and experimentation, leveraging the model's advanced natural language processing capabilities while maintaining computational efficiency. The use of Mistral 7B through Ollama facilitates robust performance in tasks such as entity tagging, relationship extraction, and text analysis on historical travelogue datasets.

By employing purely prompt-based strategies across these distinct architectures, it is possible to observe how each model's inherent language understanding influences annotation quality. This versatility ensures that the final approach is broadly applicable, offering researchers a range of deployment options without requiring them to engage in time-consuming model retraining or fine-tuning.

7.1 Technical Setup and Environment

This section details the computing resources, software libraries, and data-handling procedures used throughout the implementation. Clarifying the technical environment makes replicating or extending the methodology in future research easier.

7.1.1 Development Environment

The development environment was configured to support prompt-based annotation using modern LLMs. Key components included:

Hardware: The computational experiments for this study were conducted on a Google Colab environment. The hardware configuration included an Intel(R) Xeon(R) CPU @ 2.20GHz with 12 cores, providing ample processing power for executing large-scale natural language processing tasks. The system was equipped with 52.96 GB of total RAM, with 51.03 GB available for computational operations. An NVIDIA GPU with 22.5 GB of dedicated memory was utilized for GPU acceleration, running on driver version 550.54.15 with CUDA 12.4 support. This setup ensured efficient model inference and smooth execution of prompt-based annotation tasks without additional fine-tuning.

Operating System: Linux kernel version 6.1.85+ with an x86_64 architecture, selected for its stability and broad support for machine learning frameworks.

Software Libraries: Python 3.11.11 is selected for scripting, data manipulation, and running the model inference pipeline.

7.1.2 Dataset

The dataset for this research comprises a curated selection of 19th-century English-language travelogues sourced from open-access archives and digital libraries [Chapter 6]. These texts were chosen to capture various geographical locations, narrative styles, and historical perspectives. Each travelogue was extracted in plain text format, maintaining original punctuation and paragraph structure to preserve contextual nuances. The travelogue corpus for this study comprises 120 pages (~21,674 words) extracted from 15 historical travelogues of the Ottoman Empire.

Academic Tourism Ontology (ATO) Dataset:

This study also evaluates the effectiveness of the annotation guidelines and the performance of LLM models on the ATO datasets [51]. The ATO dataset, which consists of texts in old Dutch, is used to demonstrate how well the model can generalize to a new language and a different text genre (letters). Additionally, the preliminary annotation guidelines of this research were shared with the ATO team to support their hypothesis development process. This collaboration ensured alignment between the annotation framework and the team's research objectives, facilitating a cohesive data analysis and interpretation approach. By providing these guidelines early in the research cycle, the ATO team was able to refine their hypotheses and experimental design, ensuring that the annotation process effectively addressed the study's key questions and requirements. The ATO datasets consist of three travel narratives written in the form of letters, documenting expeditions

through Germany by students and scholars associated with the University of Groningen during the nineteenth century. The narratives, composed in Dutch, provide a rich historical and linguistic context for evaluating the capabilities of the annotation framework and the accuracy of the LLM models in processing and analyzing such texts [51]. The accounts include "Brieven, geschreven op eene wandeling door een gedeelte van Duitschland en Holland" (1809-1810) authored by Barthold Henrik Lulofs, Theodorus van Swinderen, and Jan Jacob Modderman, a manuscript dated 1816 by Reneke de Marees van Swinderen located in the Groninger Archives, and a manuscript from 1819 by Ulrich Herman Wielinga Huber held in Treseor, Leeuwarden. These texts have been digitized for accessibility. Domain specialists have identified specific letters encapsulating significant locations, individuals, items, and occurrences. Letters IV from Göttingen (July 9, 1816) and V (July 12, 1816) are exemplary samples chosen to evaluate the LLM models.

Pre-process ATO Data: The ATO team provided the gold annotations for the ATO sample dataset [51]. Subsequently, the original data was derived from these gold annotations by manually eliminating the tags and comments. To ensure compatibility with the annotation format utilized in this study, the gold annotations were manually converted to <PER> (for person entities) and <PLACE> (for location entities) tags. The resulting annotations were then systematically compared with those generated by the LLM to evaluate alignment and accuracy.

7.2 Integration of Annotation Guidelines

The annotation guidelines are represented using a structured and rule-based approach that enables the automated tagging of characters, places, and travel-related actions within travelogue excerpts.

7.2.1 Prompt Design

Each annotation is governed by predefined tagging rules, ensuring that the *main traveler* (denoted as `id=0`) and associated locations are consistently marked. The relationships between characters are explicitly encoded using attributes such as `ValidConnection` and `TravelConnection`, while spatial movements are tagged using `<Action>` elements [Section 5.2]. The system also classifies each location with attributes such as `TemporalSpan`, `IsPartOfJourney`, and `Itinerary_ID` to reflect its role in the traveler's journey. To ensure adherence to the guidelines, a multi-step validation process is implemented:

Entity Detection: LLM is prompted to identify named entities such as characters and places.

Person Attributes: Based on the relationship between the *main character* and other characters, `ValidConnection` is set to `TRUE` or `FALSE`, ensuring a valid relationship among the characters. If the `ValidConnection` is labeled as `True`, then the `TravelConnection` category (`IN GROUP`, `JOIN GROUP`, `EXIT GROUP`, etc.) is assigned according to the travel relationship between the

main character and other characters.

Person-Place Temporal Connection: If the *main character's* interaction with the place is repeated or involves a past connection, then the `TemporalSpan` is labeled as `'HABITUAL'` otherwise `'PUNCTUAL'`.

Action Tagging: Keywords indicating movement (e.g., "departed," "arrived at," "traveled to") are tagged with `<Action>` and assigned a `SpatialRelation` category (`TO`, `FROM`, `IN`, etc.). For every place entity, there should be an `<Action>` keyword that indicates the spatial relationship of the place with the *main character*.

Journey Classification: Based on spatial relations and contextual interactions, `IsPartOfJourney` is assigned either `TRUE` or `FALSE`, ensuring accurate itinerary reconstruction.

Person-Place Validation: If a place is part of the traveler's journey, it receives an `ItineraryID`, reflecting the movement sequence.

Edge Case Handling: The system cross-checks ambiguous references to interpret complex or unclear narratives.

By encoding the guidelines in this *structured* manner, the system ensures consistent, accurate, and scalable annotation of travelogues, aligning with the research objectives of capturing movement and travel-related interactions within historical narratives.

7.2.2 Annotation Data Structure

The annotation system employs an **XML-based data structure** with **inline annotations** to effectively encode the guideline rules. This approach ensures that entity tagging, spatial relations, and journey classifications are seamlessly embedded within the original text, preserving its content integrity. The XML schema is designed to maintain entity relationships, capture hierarchical dependencies, and track spatial movements throughout the narrative.

1. Inline Entity Tagging: Entities are annotated directly within the text using **inline XML tags**, explicitly marking `Person` (`<PER>`) and `Place` (`<PLACE>`) mentions.

Person (`PER`) Tagging: The *main traveler* is assigned `id=0`, while other characters receive unique IDs. Additional attributes such as `ValidConnection` and `TravelConnection` describe relationships between characters as shown below.

```
<PER id=''0''>John Stewart, Esq.</PER> was born in
<PLACE TemporalSpan=''HABITUAL'' IsPartOfJourney=''FALSE''>
```

```
London</PLACE> in the year 1749.
```

Place (`PLACE`) Tagging: Each *location* is tagged with attributes such as `TemporalSpan`, `IsPartOfJourney`, and `ItineraryID` if it is part of the traveler's movement as shown below.

```
<PER id=''0''>She</PER> <Action SpatialRelation=''FROM''>departed from
</Action> the <PLACE TemporalSpan=''PUNCTUAL'' IsPartOfJourney=''TRUE''
Itinerary_ID=''1''>Istanbul</PLACE>.
```

2. Action Annotation for Spatial Relations: *Movement-related actions* are explicitly annotated with the `<Action>` tag, capturing the `SpatialRelation` between the *main traveler* (`id=0`) and a location as shown below.

```
<PER id=''0''>He</PER> <Action SpatialRelation=''TO''>arrived at</Action>
the <PLACE TemporalSpan=''PUNCTUAL'' IsPartOfJourney=''TRUE''
Itinerary_ID=''2''>Cairo</PLACE>.
```

3. Handling Journey Sequences with `Itinerary_ID`: To accurately reflect a traveler's movement, each *visited location* is assigned a sequential `ItineraryID`, ensuring a *chronological representation* of the journey. For example:

```
<PER id=''0''>I</PER> <Action SpatialRelation=''FROM''>left</Action>
<PLACE TemporalSpan=''PUNCTUAL'' IsPartOfJourney=''TRUE'' Itinerary_ID=''1''>
Venice</PLACE> and <Action SpatialRelation=''THRU''>passed through</Action>
<PLACE TemporalSpan=''PUNCTUAL'' IsPartOfJourney=''TRUE'' Itinerary_ID=''2''>
Vienna</PLACE> before <Action SpatialRelation=''TO''>reaching</Action>
<PLACE TemporalSpan=''PUNCTUAL'' IsPartOfJourney=''TRUE'' Itinerary_ID=''3''>
Prague</PLACE>.
```

This ensures that the *traveler's itinerary* is reconstructed accurately, following the structured guidelines.

4. Relationship Encoding Between Characters: The system captures *connections between the primary character and other characters* using attributes within the `<PER>` tag: `ValidConnection="TRUE"` indicates a confirmed relationship and `TravelConnection` defines the nature of the connection (`IN_GROUP`, `JOIN_GROUP`, `EXIT_GROUP`, `NO_REL`). For example:

```
<PER id=''0''>I</PER> traveled with <PER id=''1'' ValidConnection=''TRUE''
TravelConnection=''IN\_GROUP''>Maya</PER> from <PLACE TemporalSpan=''PUN-
```

```
CTUAL'' IsPartOfJourney=''TRUE'' Itinerary_ID=''1''>Paris</PLACE>.
```

By implementing this **XML-based inline annotation structure**, the system maintains data consistency, facilitates easy *post-processing*, and ensures that travel narratives remain intact while capturing detailed annotations.

7.2.3 Handling ambiguous narratives

The annotation guidelines include cases where edge cases and ambiguities must be resolved so the system can interpret complex or unclear narratives. Key strategies include:

1. Managing Temporal Ambiguities: If a place is referenced in the past or as part of a background story but is not part of the current narrative, then IsPartOfJourney is set to False.

e.g.: "*Two years* ago, I visited Constantinople."

The narrative is indicating a past journey, so IsPartOfJourney=FALSE.

2. Unclear Spatial Relationships: When the spatial relationship between a character and a place is unclear or does not exist, the SpatialRelation labeled NO_REL and IsPartOfJourney is set to False. So, only relevant places will be included in the itinerary.

3. Future Intentions Phrases indicating future travel plans, such as "will travel to" or "plan to visit," are identified, and associated destinations are extracted and excluded from the itinerary (IsPartOfJourney=False).

4. Unclear Character-Place Associations: If a place is mentioned but not directly tied to the primary character (id=0), the IsPartOfJourney attribute is set to FALSE.

5. Managing Multiple Characters: The system manages interactions among various characters by assigning distinct IDs and defining their relationships (e.g., ValidConnection, TravelConnection). This guarantees an accurate representation of the primary character's journey.

6. Addressing Character Movement Over Multiple Locations: If the primary character travels through several places before reaching a final destination, the system enumerates a detailed itinerary by sequencing places based on their spatial relationships (FROM $\rightarrow$ THRU\NEAR $\rightarrow$ TO)

By implementing these handling mechanisms, the system ensures that the annotations are contextually accurate, avoid false positives, and maintain fidelity to the travelogue's narrative flow.

7.3 OpenAI GPT-Based Annotation Module

This study initially leveraged the OpenAI platform to evaluate the proposed hypothesis, utilizing ChatGPT with structured prompts to provide instruction-driven annotations. Building upon this foundation, a domain-specialized GPT-4 model, referred to as the 'Travelogue Annotator' version-1, was developed using state-of-the-art LLM [52]. This model was configured through OpenAI's MyGPT function, incorporating a systematically designed system prompt to enhance its performance in annotating historical travel narratives.

The preliminary annotation guidelines and results were presented in selected workshops, including the *DHd-2024 Passau: Workshop on Generative KI, LLMs, and GPT in Digital Editions* and the *Regensburg-2024: Workshop on Interlinking Text and Data with Semantic Annotation and Ontology Design Patterns to Analyze Historical Texts* [53]. These discussions provided valuable insights into the annotation model's effectiveness and adaptability across different historical datasets.

For the final implementation, an OpenAI *GPT-based annotation assistant* was systematically developed to process and analyze travelogue excerpts. This assistant was designed to accurately identify and annotate connections between characters and places, incorporating structured entity tagging, spatial relationship detection, and journey segmentation. Integrating domain-specific instructions, two/few-shot learning, and post-processing validation ensured the model's robustness in handling complex historical texts while maintaining annotation consistency.

This section details the methodology and implementation of an OpenAI GPT-based annotation module, focusing on its application in travelogue analysis. The module leverages prompt engineering, in-context learning, and structured entity tagging to annotate historical travel narratives accurately. The assistant is designed using OpenAI GPT-based models, configured for domain-specific tasks through carefully crafted system instructions, API settings, and structured input-output processing.

7.3.1 Methodology for Using OpenAI GPT Models

The annotation module's effectiveness heavily depends on prompt design, input structuring, and output parsing. The methodology follows a three-step process:

1. **Prompt Design:** The assistant is configured with domain-specific system instructions that define its role in analyzing and annotating connections between characters and places within travelogues. The prompt includes: **(i)** A clear task definition specifying the annotation objective, **(ii)** Examples of correctly annotated travelogue excerpts to guide in-context learning, and **(iii)** Explicit rules for identifying entity relationships, spatial movements, and character-place associations.

2. **Input Formatting:** To maintain consistency, raw travelogue texts are preprocessed to ensure proper sentence segmentation and formatting. This enables the model to recognize patterns

effectively.

3. **Output Parsing:** The assistant generates structured annotations based on entity tagging conventions. The output is validated against predefined annotation rules to ensure compliance and consistency.

7.3.2 Few-Shot Learning Approach

To improve annotation accuracy, the model is configured with a **two-shot and few-shot learning approach**. This involves providing a small set of annotated examples within the prompt to enhance its ability to generalize across different travelogue excerpts. This approach offers several key advantages, particularly in reducing manual effort, enhancing contextual adaptability, and improving annotation consistency. One of its primary benefits is **minimized manual supervision**, as the model leverages a few high-quality annotations to guide its predictions, thereby reducing the need for extensive manual corrections. Additionally, it demonstrates **contextual adaptability** by adjusting to domain-specific linguistic patterns found in historical texts, ensuring more accurate entity recognition within specialized datasets. Another critical advantage is **improved consistency**, as the model adheres to established annotation guidelines, minimizing variability in entity recognition and ensuring uniformity across annotations. Figure 7.2 illustrates the difference between standard inference and few-shot learning.

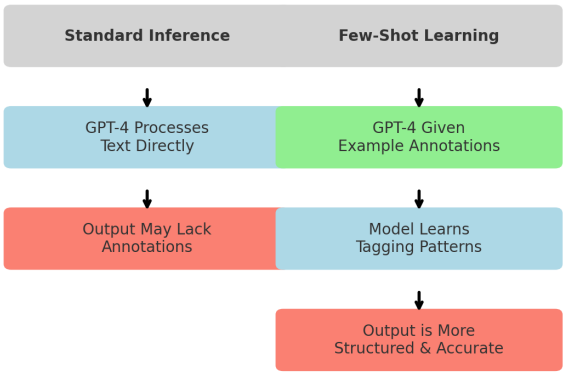


Figure 7.2: Illustration of standard vs. few-shot inference in OpenAI GPT-based annotation.

7.3.3 Implementation Details

The OpenAI GPT-based annotation module is implemented using GPT-4o, 4-turbo, and 3.5-turbo API. The following illustrates the configuration settings for an OpenAI GPT-based annotation assistant designed to process and analyze travelogue excerpts. This assistant utilizes in-context

learning and domain specialization prompts to accurately annotate historical texts by identifying relationships between characters and places. To optimize annotation performance, the system is configured with specific API parameters. The key components relevant to the functionality and adaptability of the assistant are described, and how they contribute to the overall annotation workflow is explained as follows:

1. **Name Field:** The Name field identifies the specific assistant configured for travelogue annotation tasks. This naming convention helps distinguish between various assistants, especially when multiple models are tailored for different tasks or datasets.

2. **Assistant ID:** A unique identifier assigned to the assistant. This ID ensures the assistant's precise tracking and management across multiple annotation tasks. It allows for consistent API calls and logging of annotation outputs, which is critical for version control and reproducibility in large-scale annotation projects.

3. **System Instructions:** Contains a *domain-specific prompt* that instructs the OpenAI GPT model on how to perform the annotation task. The system instructions explicitly define the scope of the assistant's role. For example, it specifies that the assistant analyzes connections between characters and places within travelogue excerpts, highlights relationships, and identifies spatial dynamics. Furthermore, this prompt is a foundation for *in-context learning* by providing clear guidelines and example tasks. The model leverages these instructions to understand the nature of the annotation and adapt its responses accordingly. The prompt ensures that the assistant applies domain knowledge effectively, distinguishing between different types of entity relationships.

4. **Model Selection:** The selected model (e.g., gpt-4o) determines the capability and performance of the assistant. Such as GPT-4 is known for its advanced language understanding and contextual reasoning abilities. It is well-suited for complex annotation tasks requiring nuanced comprehension of historical texts and travel narratives. The model's ability to process and infer context enhances the accuracy of *in-context learning* and annotation consistency.

5. **Model Configuration:** The model is configured in two distinct modes, namely:

   Response Format: The response format is set to "text," ensuring that the outputs are generated in a readable and structured format suitable for annotation.

   Hyperparameters:

   - `Temperature` = 0.0 (to ensure deterministic output)
   - `Max Tokens` = 128,000 (to capture long travelogue excerpts)
   - `Top-P` = 1 (for balanced response diversity)

The overall annotation pipeline of the OpenAI Assistant GPT model is described below, detailing the steps involved in creating and refining the annotated output. This workflow ensures that the

assistant can generate accurate, contextually relevant, and user-aligned responses.

1. **Input and Prompting:** The assistant receives raw travelogue text (text inside `'<Data>'` tag; Section 7.7) as input. The system instructions provide context for interpreting these texts, focusing on character-place relationships and spatial dynamics.
2. **API Calls:** Requests are made via OpenAI's API, with formatted prompts sent in a structured text format.
3. **In-Context Learning:** The assistant leverages example annotations provided within the prompt (not shown in the figure 7.3 but implied by the system instructions) to understand the desired output format and annotation conventions. This in-context learning enables the model to generalize from a few examples and apply similar patterns to new data.
4. **Processing and Annotation:** Using the OpenAI GPT-based model, the assistant analyzes the text, identifies entities (characters and places), and tags relationships based on the specified guidelines.
5. **Output and Validation:** The model generates structured annotations in XML-based format, which human annotators can review and refine as needed. The deterministic output ensures consistency across annotations.
6. **Adaptability Across Datasets:** The configuration allows easy adaptation to different datasets or annotation tasks. By adjusting the system instructions and prompt content, the assistant can be tailored to various domains while leveraging the same underlying OpenAI GPT-based model capabilities.

7.3.4 Key Challenges and Solutions

While implementing the OpenAI GPT-based annotation module, several challenges were encountered. Table 7.1 presents these challenges along with the corresponding solutions.

Table 7.1: Key challenges and solutions in OpenAI GPT-based annotation.

Challenge	Solution
Misclassification of entities	Enhanced prompt clarity and few-shot examples to reinforce correct tagging.
Inconsistent spatial relation tagging	Introduced explicit formatting in the prompt to highlight movement actions and relationships.
Handling ambiguous travel references	Implemented post-processing validation rules to filter out hypothetical and past journey mentions.
Ensuring XML-compliant output	Added sample output structure to reinforce correct malformed annotations.

By addressing these challenges, the annotation module ensures reliable and accurate extraction of travel-related entities and spatial movements within historical texts.

The OpenAI GPT-based annotation module leverages structured prompts, two/few-shot learning, and systematic post-processing to annotate historical travel narratives accurately. The methodology ensures adaptability across diverse datasets while minimizing manual intervention. By refining the model's input-output design and addressing key challenges, the proposed system enhances entity recognition and spatial reasoning within travelogues, contributing to high-quality *gold standard* annotations.

7.4 LLaMA-Based Annotation Module

This section describes the implementation of the *LLaMA-based annotation module* utilized through the *Ollama platform*. Following a two-shot learning approach, the model leverages a structured inference pipeline to process and annotate travelogue excerpts. Key differences from OpenAI GPT-based annotation models, including prompt template customization, inference methodology, model-specific optimizations, and observed limitations, are highlighted.

7.4.1 LLaMA Model and Integration Process

The annotation module is implemented using the *LLaMA 3 8B instruct model (fp16)* through the *Ollama platform*. The LLaMA model is selected for its open-source accessibility, efficient inference speed, and fine-tunable capabilities, making it suitable for travelogue annotation tasks. Unlike OpenAI GPT-based solutions, which depend on API-based interactions, Ollama facilitates local model deployment through a custom model file. It employs a comparable methodology for *Prompt Design*, *Input Formatting*, and *Output Parsing*, as outlined in Section 7.3.1. This approach ensures cost-effective annotation processing, data privacy, and domain adaptation.

7.4.2 Two-Shot Learning Approach

The *two-shot learning approach* is employed to enhance annotation accuracy, leveraging a small number of examples to guide the model's output due to its limited content size. Unlike traditional fine-tuning, which requires extensive labeled data, two-shot learning allows LLaMA to generalize annotation rules from a minimal dataset. The two-shot approach is designed to enhance model performance by leveraging minimal but effective examples for annotation learning. It relies on **example pairs**, where the model is provided with two annotated travelogue excerpts that illustrate proper entity tagging, spatial relationship classification, and travel event detection. This approach is further reinforced through **prompt conditioning**, in which a structured template ensures that the model correctly associates user-provided travelogue text with the expected annotation format. As a result, the model demonstrates **generalization** by applying learned annotation patterns to unseen text excerpts after processing just two examples, thereby reducing the need for extensive manual corrections and improving annotation efficiency.

7.4.3 Implementation Details

The *LLaMA 3 8B instruct model's model file* is configured to specialize the LLaMA model for historical travelogue annotation. The *'Modelfile'* consists of the following components:

1. **Model Deployment:** The *LLaMA 3 8B instruct model* is loaded within Ollama, ensuring an optimized runtime environment with fp16 precision for faster inference.
2. **Template-Based Prompt Structuring:** A well-defined system prompt is implemented using structured delimiters such as `|start_header_id|` and `|end_header_id|` to segment system instructions, user input, and assistant responses.
3. **System-Level Instructions:** The *SYSTEM* section embeds task-specific guidelines and annotation rules within the model, functioning similarly to OpenAI GPT assistant system prompts.
4. **Parameter Configuration:** The model parameters, including context length, temperature, and stop sequences, are configured to enhance annotation consistency.
   - The temperature is set to *0.00*, ensuring deterministic outputs that reduce annotation variability.
   - The context window is set to *8192* tokens, though still limited for multi-page processing.
   - The num_keep is set to *2500*, allowing partial retention of prior context in longer interactions.
5. **Stop Sequences for Output Control:** The stop parameters include:
   - `|start_header_id|`
   - `|end_header_id|`
   - `<eot_id>`

   These constraints prevent overgeneration and ensure that responses remain structured and within the required format.

7.4.4 Differences in Approach from OpenAI GPT Model

The LLaMA-based annotation module differs significantly from OpenAI GPT-based approaches in terms of *prompt structuring, inference pipeline, and output parsing*.

Prompt Template and Structuring

The OpenAI GPT-based model (as discussed in Section 7.3) primarily relies on a *free-text prompt structure* with *instruction-based conditioning*. In contrast, the LLaMA module uses *structured prompt templates* with predefined delimiters.

LLaMA Prompt Structure:

```
|start_header_id|system|end_header_id|
<System Instruction>
|eot_id|
|start_header_id|user|end_header_id|
<User Input>
|eot_id|
|start_header_id|assistant|end_header_id|
<Model Response>
|eot_id|
```

OpenAI GPT Models Prompt Structure:

```
System: ''Provide annotations for the given travelogue excerpt...''
User: ''<Travelogue Excerpt>''
Assistant: ''<Annotated Output>''
```

The structured formatting in LLaMA ensures modular parsing and controlled response generation, making annotation tasks more predictable.

Inference Pipeline Differences

LLaMA's inference pipeline follows a *sequential structured processing* approach, whereas OpenAI GPT models typically rely on a *flexible, open-ended response generation approach*. Key differences are addressed in Table 7.2, highlighting the distinctions between the methods discussed in the study. This table provides a clear comparison to facilitate a deeper understanding of the different LLM models.

Table 7.2: Comparison of LLaMA (Ollama) vs. OpenAI GPT-Based Annotation

Aspect	**LLaMA (Ollama)**	**OpenAI GPT Model (API-Based)**
Prompt Structure	Uses header delimiters for structured I/O	Free-text prompting with flexible response
Model Hosting	Runs locally on a machine (no API cost)	Cloud-based inference (requires API access)
Context Window	Limited (8192 tokens), struggles with long texts	Typically larger (128,000 tokens in GPT-4o and GPT-4 Turbo)
Entity Recall	Weaker for rare historical places	Stronger general knowledge base

7.4.5 Observed Strengths and Limitations

Based on experimental results, the LLaMA-based annotation module demonstrates several advantages and limitations compared to OpenAI GPT-based annotation models.

Strengths

The advantages of the LLaMA-based annotation module are as follows:

- **Open-Source Accessibility:** LLaMA is a fully open-source model, allowing it to be used locally without API costs, making it highly cost-effective compared to proprietary models like OpenAI GPT.
- **Local Implementation:** Unlike OpenAI GPT-based models, which require cloud-based API access, LLaMA can be deployed entirely offline, making it suitable for secure environments with data privacy concerns.
- **Structured Prompt Handling:** The use of header delimiters ensures that input-output parsing remains controlled and modular.
- **Optimized Memory Usage:** The LLaMA model provides faster inference with *fp16 precision*, making it computationally efficient for large-scale annotation projects.
- **Greater Customization Control:** Parameters like context length retention and stop sequences allow fine-tuned inference control.

Limitations

Despite its advantages, the LLaMA-based annotation module exhibited several challenges and limitations:

- **Less Flexible than OpenAI GPT Models:** The strict formatting reduces *free-text adaptability*, making it less effective for *complex contextual reasoning tasks*.
- **Potential Underfitting in Two-Shot Learning:** While effective in simple annotation tasks, the model struggles with edge cases that require deeper contextual understanding.
- **Struggles with Complex Reasoning Tasks:** The LLaMA model showed difficulty following multi-step instructions and handling complex travelogue annotations, as it requires more explicit fine-tuning to improve domain-specific accuracy.
- **Knowledge Cut-Off and Compressed Size:** Due to limited complexity and knowledge cut-off, it struggled to identify historical place entities that are uncommon or no longer in use. This led to gaps in annotation accuracy for such entities.
- **Limited Context Window:** The model exhibited an inability to maintain the specified output structure when processing multiple pages due to its constrained context length. This issue

affected long-form annotation consistency.

- **Text Modifications and Omissions:** Despite explicit instructions to preserve the original data, the model occasionally altered or omitted portions of the input text, leading to potential data integrity issues.

The LLaMA-based annotation module provides an efficient and structured approach for annotating travelogue excerpts, offering advantages in output consistency, deterministic response generation, and modular prompt handling. However, compared to OpenAI GPT-based models, it struggles with complex multi-step reasoning tasks, entity recognition for uncommon historical places, and long-context coherence.

By leveraging open-source accessibility, cost-effective local deployment, and structured processing, the LLaMA-based system is particularly suited for controlled annotation tasks. However, further fine-tuning and integrating additional function-calling mechanisms for NER may be necessary to address highly complex reasoning tasks, process lengthy texts, and improve the recognition of extensive entities.

7.5 Mistral-Based Annotation Module

This section introduces the *Mistral-based annotation module*, which utilizes the `mistral:7b-instruct-q8_0` model through the *Ollama platform* to enable structured annotation of travelogues. The implementation adopts a custom *'Modelfile'* approach, similar to that used in LLaMA, allowing for local execution, deterministic inference, and domain-specific customization [Section 7.4]. By incorporating *structured input formatting* and *controlled response generation*, this module is designed to improve annotation accuracy while maintaining privacy and computational efficiency. Furthermore, this section explores key methodological aspects, such as prompt design, implementation parameters, and a comparative analysis with OpenAI *GPT-based annotation models*.

7.5.1 Mistral Model and Integration Process

The Mistral-based annotation module is implemented through the *Ollama platform*, facilitating on-device execution and eliminating reliance on cloud-based inference models. This approach leverages *Modelfile-based integration* to establish a structured and systematic annotation workflow, ensuring consistency and interpretability across historical travelogue datasets. The methodology aligns with established practices in *Prompt Design*, *Input Formatting*, and *Output Parsing*, as detailed in Section 7.3.1, thereby maintaining methodological rigor and reproducibility in the annotation process.

The Mistral-based approach offers several key advantages, particularly in terms of computational efficiency, privacy, and structured annotation control. One of its primary benefits is **computational**

efficiency, as the *local execution* of Mistral significantly reduces *operational costs*, making it a viable alternative to API-dependent architectures. Additionally, the approach ensures **privacy-preserving annotation** by keeping all computations strictly *on-device*, thereby maintaining *data confidentiality*, which is particularly crucial when processing sensitive historical texts. Another advantage is **structured output control**, achieved through the incorporation of *predefined stop sequences and instruction delimiters* within the model file. This mechanism enforces output constraints, enhancing *response predictability* and ensuring *annotation coherence*, making it a robust solution for structured text processing.

7.5.2 Two-Shot Learning Approach

The two-shot learning approach improves annotation accuracy by leveraging a minimal set of examples to guide the model's output, which is particularly suited for scenarios with limited content size. This method aligns closely with the implementation employed in LLaMA, as detailed in Section 7.4.2, ensuring consistency in methodology and facilitating comparative analysis. Incorporating carefully selected examples enhances the model's ability to generalize and produce contextually relevant annotations, even with constrained input data.

7.5.3 Implementation Details

The Mistral-based annotation module is governed by a *structured model file configuration*, which ensures *controlled inference*, *structured text generation*, and *tool-based extensions*. The subsequent subsections provide a detailed description of the *prompt structuring*, *model parameters*, and *response control mechanisms* integrated into the model file. The 'Modelfile' comprises the following key components:

1. **Model Deployment:** The `Mistral 7b instruct model` is deployed within Ollama, utilizing an optimized runtime environment configured with *q8_0 precision* to enhance computational efficiency and accelerate inference speeds. This deployment approach ensures high performance while maintaining resource efficiency, making it well-suited for scalable and precise annotation tasks.

2. **Template-Based Prompt Structuring:** The Mistral-based annotation module incorporates instruction-based segmentation to ensure structured and interpretable input-output exchanges. The model processes annotations through delimited instruction blocks, utilizing specific markers such as `[INST]`, `[/INST]`, and `</s>` to delineate system, user, and assistant interactions clearly. This approach enforces a well-defined communication flow, enhancing the clarity and consistency of the annotation process. By segmenting instructions and responses in this manner, the module maintains robust control over text generation, ensuring that outputs align with predefined guidelines and facilitating more accurate and reliable annotations.

3. **System-Level Instructions:** The SYSTEM section integrates task-specific guidelines and annotation rules into the model, operating in a manner analogous to LLaMa or OpenAI GPT assistant system prompts. The system prompt is injected *only once* at the start of the session to define annotation rules and ensure task consistency throughout the interaction. This is achieved through the following structured format:

   ```
   [INST] {{ $.System }} [/INST]
   ```

   This approach ensures the model adheres to the predefined guidelines, maintaining coherence and accuracy in the annotation process while minimizing redundancy. The module establishes a clear framework for subsequent interactions by embedding instructions at the session's outset, enhancing efficiency and reliability.

4. **Parameter Optimization:** The model parameters, including context length, temperature, and stop sequences, are configured to enhance annotation consistency.
   - Temperature set to *0.00* for deterministic outputs.
   - Extended context window (*32768 tokens*) for multi-page annotation capability.
   - Memory retention (*5250 tokens*) to maintain annotation continuity.

5. **Predefined Stop Sequences:** The stop parameters include:
   - `[INST]`
   - `[/INST]`
   - `</s>`

   These constraints prevent excessive response generation and ensure that responses remain structured and within the required format.

Mistral Prompt Format

The Mistral model enforces *structured instruction templates* similar to LLaMA. The following structure is employed to regulate user queries and assistant responses:

```
[INST] <System Instruction> [/INST]
[INST] <User Input: Travelogue Excerpt> [/INST]
<Assistant Response>
</s>
```

7.5.4 Differences in Approach from OpenAI GPT Models

The Mistral-based annotation system differs significantly from OpenAI GPT-based approaches in terms of prompt structuring, inference pipeline, and output parsing:

Table 7.3: Comparison of Mistral vs. OpenAI GPT-Based Annotation

Aspect	Mistral (Ollama)	OpenAI GPT Models (API-Based)
Prompt Format	Uses `[INST]` and stop sequences	Free-text prompting
Model Hosting	Local execution (no API cost)	Cloud-based inference (requires API)
Context Window	32768 tokens (better for multi-page texts)	Up to 128K (GPT-4o,4 Turbo)
Historical Entity Recall	Struggles with rare place names	Strongest general knowledge base

7.5.5 Observed Strengths and Limitations

Based on experimental results, the Mistral-based annotation module exhibits several advantages and limitations compared to **LLaMa**-based annotation models.

Strengths

The advantages of the Mistral-based annotation module are as follows:

- **Structured Prompt Handling:** Using delimited instruction blocks ensures that input-output parsing remains controlled and modular.
- **Offline Deployment:** Running entirely on local hardware eliminates the need for third-party API services, ensuring privacy compliance.
- **Custom Model File Adaptability:** The model can be extended through external tools, enabling further domain specialization.
- **Context Window Size:** Larger context window size compared to the LLaMa model.

Limitations

Despite having a larger context size compared to **LLaMa**-based annotation models, the Mistral-based annotation module faced similar challenges and limitations as LLaMa.

- **Limited Generalization for Complex Reasoning Tasks:** The model struggles with multi-step annotation procedures that require implicit knowledge.
- **Inconsistencies in Historical Entity Recognition:** Due to pretraining limitations, Mistral fails to tag lesser-known historical places accurately.
- **Potential Data Alterations:** Despite explicit instructions, the model occasionally modifies input text, affecting annotation reliability.

The Mistral-based annotation module provides a structured, cost-effective alternative to OpenAI GPT-based annotation systems. Compared to LLaMA, it offers a more significant context window, improving multi-page annotation consistency. However, like LLaMA, it struggles with historical place recognition, instruction-following consistency, and annotation coherence.

Similar approaches, as outlined in Section 7.4.5, can be employed further to enhance the capabilities of the Mistral-based annotation module. For instance, fine-tuning the model on domain-specific datasets and integrating advanced function-calling mechanisms for NER could significantly improve its ability to handle highly complex reasoning tasks, process lengthy texts, and enhance the recognition of extensive entities. These strategies would enable the model to better adapt to specialized annotation requirements, increasing its versatility and accuracy in addressing challenging use cases. Such enhancements align with the broader goal of optimizing annotation workflows while maintaining computational efficiency and task-specific precision.

7.6 Inference Pipeline and Automation

This section outlines the inference pipeline and automation process in the annotation workflow. The annotation is performed book-wise, with each page of the raw travelogue text provided as input to the LLM models for annotation. The raw travelogue text is extracted from post-processed documents, encapsulated within `'<Data>'` tags [Section 6.5]. The LLMs respond with annotated data using XML tags, which are then integrated into the original document, ultimately producing an annotated version of the entire travelogue.

7.6.1 Overall Workflow from Input Text to Annotated Output

The overall annotation workflow, from the initial input of raw text to the generation of annotated output, is depicted in Figure 7.3. The steps involved in this process are as follows:

1. **Create Assistant:** The annotation process begins by creating a new assistant instance responsible for processing user prompts and generating annotations.
2. **Create Thread:** A new annotation thread is initialized to manage a specific book or section of travelogue text.
3. **Start Thread:** The thread is activated, setting up the context for processing page-wise annotation requests.
4. **Raw Travelogue Text Input:** The unannotated travelogue text, extracted from post-processed documents (Per Book), is fed into the workflow in a page-wise manner (Per page data at a time).
5. **User Prompt with Data:** The LLM models receive prompts containing contextual information about the raw travelogue text, guiding the annotation process.

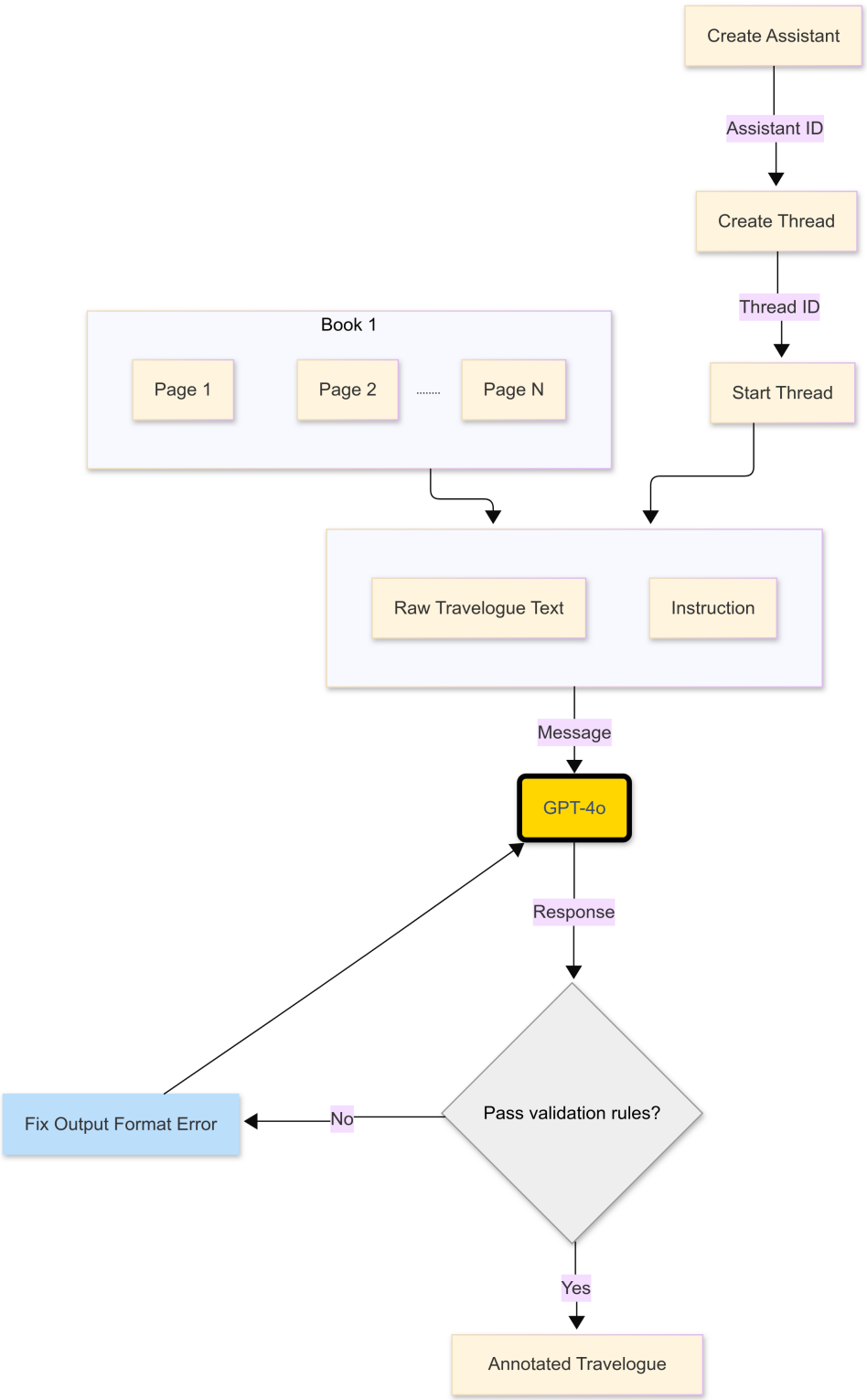


Figure 7.3: Overall Workflow for Annotating Travelogues Using LLM Models

6. **LLM Processing:** The LLM models perform the following operations:
    - **API Calls:** The assistant sends prompts via API calls to the LLM for processing.
    - **Text Processing and Annotation:** The LLM identifies entities, spatial relationships, and travel connections, annotating the text with XML tags.
    - **Output Parsing and Validation:** The annotated output is parsed, validated, and checked for compliance with annotation guidelines.
7. **Annotated Travelogue Integration:** The XML-annotated output replaces the original text in the document, and the annotated travelogue is compiled book-wise to produce the final annotated output.

7.6.2 Error Handling Strategies

The following error handling and fallback strategies are implemented to ensure the robustness of the annotation process :

Input Validation: Before submitting text to the LLM, the input is validated to ensure it conforms to the expected format. This prevents processing errors due to malformed or incomplete inputs.

Annotation Integrity Checks: The annotated output is parsed and checked manually for compliance with annotation guidelines. The process retries annotation with modified prompts if inconsistencies or formatting errors are detected.

The inference pipeline described in this section facilitates the automated annotation of historical travelogues using LLM models. The pipeline ensures accurate, efficient, consistent annotation output by leveraging structured input prompts and error-handling mechanisms. The annotated travelogues produced through this process serve as valuable resources for further analysis and research in historical travel literature.

7.7 Post-possessing Data:

To evaluate the annotation data generated by the LLM models, post-processing is required to align it with the format specified by the *Nervaluate* Python module, as outlined in chapter 8 [54]. This step ensures compatibility with the evaluation framework and accurately assesses the model's performance against the defined metrics. Below is a step-by-step explanation of its functionality:

Tag Replacement and Information Extraction

Regular expressions are employed to identify and process annotated tags within the input text. Each match is analyzed to extract the tag name, attributes, and inner text while ensuring that entities' adjusted start and end positions are accurately maintained after removing tags. The extracted

attributes are then parsed into a structured dictionary, where their values are normalized to ensure consistency. These details are accumulated into a structured list, referred to as 'tag_info_list', which contains the tag name, attributes, start and end positions, and inner text. Finally, the annotated tags within the input text are replaced with their corresponding inner text, eliminating unnecessary spaces and preserving the original text structure while returning both the cleaned text and the extracted tag information.

NER Data Preparation

The system processes each tag and generates structured NER data in a format suitable for evaluation. It efficiently handles different tag types, such as 'PER', 'Action', and 'PLACE', by mapping attributes (e.g., 'id', 'SpatialRelation', and 'TemporalSpan') to meaningful labels. For each tag, multiple labels are created to represent various relationships or attributes, such as "ValidConnection" and "TravelConnection" for 'PER', or "TemporalSpan" for 'PLACE'. To ensure consistency, unique label definitions are aggregated in a 'tags_set', allowing for structured reference during processing. Finally, all extracted information is compiled into a 'processed_data' list, where each entry consists of a dictionary containing the assigned labels, entity spans, and corresponding text segments, facilitating seamless evaluation and structured entity representation.

Handling of Specific Entity Types

The annotation system processes three primary entity types: 'PER' (Person), 'Action', and 'PLACE'. For person entities ('PER'), it extracts attributes such as 'id', 'ValidConnection', and 'TravelConnection', assigning labels that incorporate both the 'id' and relationship attributes to ensure proper entity linkage. Action tags ('Action') are processed by capturing the 'SpatialRelation' attribute and generating a corresponding label representing the spatial interaction between entities. Place entities ('PLACE') are handled by extracting attributes like 'TemporalSpan' and 'IsPartOfJourney', with multiple labels assigned to each 'PLACE' entity to represent its temporal and journey-related characteristics. This structured labeling approach ensures accurate representation of relationships and interactions between characters, actions, and locations within the annotated text.

Output

The function returns 'processed_data', a list of structured annotations ready for evaluation. The 'tags_set' contains all unique labels generated, providing a reference for the evaluation framework.

This process ensures accurate preprocessing of annotated data by cleaning the text and extracting essential details. It maintains compatibility with the **Nervaluate** framework for structured evaluation, allowing seamless integration. Additionally, it offers flexibility to handle multiple entity types and attributes, making it well-suited for domain-specific NER tasks.

7.8 Summary

This chapter outlined the design and implementation of the *automated annotation pipeline* for historical travelogues using LLM-based models. The system successfully automates the annotation process by leveraging *structured input formatting, scalable inference mechanisms, and robust validation techniques*. Through the integration of OpenAI **GPT-based models, the quantized version of LLaMA, and Mistral models**, alongside *structured model files* and *systematic prompt engineering*, the pipeline achieves consistent entity recognition, travel connection classification, and spatial annotation. Additionally, comprehensive *error handling mechanisms*, such as API request monitoring and structured output parsing, were implemented to improve annotation reliability. Furthermore, a *post-processing module* was developed to align the *annotated XML data* with the **Nervaluate evaluation framework**, ensuring compatibility with established assessment metrics.

Despite these advancements, several challenges remain. One key issue is the *inconsistent instruction-following capabilities* of quantized LLM models, particularly with **LLaMA and Mistral**, which struggle with *complex multi-step reasoning tasks*, impacting annotation accuracy. Additionally, the *limited availability of historical place name databases* results in *occasional errors or omissions* in entity recognition. The *finite context window of LLMs* poses another challenge, as longer travelogue excerpts may be *truncated or misaligned*, requiring manual intervention or alternative input segmentation strategies. While *Nervaluate integration* ensures compatibility with evaluation metrics, additional *custom preprocessing scripts* are needed to handle *structural inconsistencies in XML output*. Moreover, the computational overhead associated with large-scale inference presents efficiency challenges, highlighting the need for further *optimization strategies* to reduce annotation runtime and improve throughput.

To formally assess the quality of annotations, an *evaluation framework* has been established using **Nervaluate**, a Python module designed for NER performance assessment [54]. The annotated data undergoes *post-processing* to align with the expected evaluation format, allowing for *precision, recall, and F1-score calculations*. The assessment also incorporates *span-based and token-based evaluations*, measuring the correctness of *nested or overlapping entity spans* within complex travel descriptions. Additionally, comparative analysis across OpenAI *GPT, LLaMA, and Mistral* models will identify the strengths and weaknesses of different architectures in travelogue annotation. Error categorization, focusing on *incorrect tagging, missing entities, and spatial relation misclassifications*, will provide deeper insights into areas requiring refinement. The results of this evaluation will be discussed in the next chapter [Chapter 8], where a detailed breakdown of model performance will highlight key findings and inform future improvements in *prompt design, model selection, and annotation consistency*. The following chapter will present the *formal evaluation of annotation performance*, offering *quantitative insights* into the effectiveness of *LLM-driven travelogue annotation*.

Chapter 8

Evaluation

This chapter presents a comprehensive discussion of the data evaluation framework, results, and their analysis. The evaluation framework is meticulously designed to assess the performance of the proposed methodology, incorporating metrics such as *precision, recall, and F1 score* to quantify the effectiveness of the Named Entity Recognition (NER) system [Section 4.6]. The results are systematically analyzed to identify patterns, strengths, and limitations, providing insights into the model's ability to extract and classify entities within the dataset accurately. Furthermore, the analysis highlights key findings, including the impact of specific annotation guidelines and the influence of dataset characteristics on model performance. This rigorous evaluation validates the proposed approach and offers valuable directions for future research and optimization in the historical travelogue information extraction and itinerary reconstruction.

8.1 Evaluation Framework

This task focuses on extracting named entities, precisely *person* and *place* entities and *action tags* from text. Extracting *person* and *place* entities provides foundational information for constructing contextual frameworks, while *action tags* capture the dynamic behaviors and events associated with these entities. In this context, seven relationship types are defined to capture the nuances of these interactions:

a. **Valid Relation:** Accuracy of the tagged named entity.

b. **Spatial Relation:** The spatial relationship between person and place entities.

c. **Temporal Span:** Past or specific temporal relationships between person and place entities.

d. **IsPartOfJourney:** Travel-related relationships between person and place entities.

e. **Valid Connection:** Relationships between person and person entities.

f. **Travel Connection:** Travel-related relationships between person and person entities.

For further details on these relationships, please refer to the proposed annotation guidelines [Chapter 5].

The evaluation process leverages the *Nervaluate* Python module, which follows the evaluation methodology outlined in the *SemEval 2013 - 9.1 task* [55] [54]. This methodology aligns well with the task's objectives and provides robust metrics for assessing model performance. *Nervaluate* 's framework is specifically designed to address the nuanced requirements of NER tasks, enabling a comprehensive and systematic evaluation of Large Language Models (LLM).

The *Travelogue Corpus*, comprising 120 pages (~21,674 words) extracted from 15 historical travelogues of the Ottoman Empire written in English, serves as the dataset for this task [Section 6.4]. The gold annotations for this corpus are hidden. Participating LLM models are tasked with reproducing these annotations by identifying entities and their associated class information. The required output format for the models is as follows:

```
IdSentence|startOffset-endOffset|text|type
```

Each recognized entity must be represented on a single line, specifying:

- The ID of the sentence containing the entity.
- The entity's starting and ending character positions within the sentence.
- The text of the entity.
- The entity type.

If multiple entities are identified within the same sentence, each must appear on a separate line. This standardized output format ensures consistency and objectively compares model performance.

However, The *nervaluate* library supports both formats shown below, which represent the same Named Entity Recognition (NER) data but use different approaches to encode the information.

- **Prodigy-Style Spans Format:**
  - Uses a list of dictionaries to specify entities' label, start, and end positions.
  - Example:

```
true = [
    [{'label': 'PER', 'start': 2, 'end': 4}],
    [{'label': 'LOC', 'start': 1, 'end': 2},
     {'label': 'LOC', 'start': 3, 'end': 4}]
]
```

- **IOB/CONLL Format:**
  - Uses a sequential tagging system for each token.

- Example:

```
true = [
    ['O', 'O', 'B-PER', 'I-PER', 'O'],
    ['O', 'B-LOC', 'I-LOC', 'B-LOC', 'I-LOC', 'O']
]
```

This study adopts the *Prodigy-style spans* format for several compelling reasons. First, this format explicitly specifies the entity type and its precise position within the text, enhancing both clarity and precision in entity representation. Second, it is particularly effective for handling overlapping or nested entities, which are frequently encountered in complex textual data. Third, the *Prodigy-style spans* format efficiently stores information only about actual entities rather than every token, thereby reducing redundancy and improving readability. This efficiency is especially advantageous for processing longer texts, such as those in the *Travelogue Corpus*, where entities are often sparsely distributed across extensive passages. While both *Prodigy-style spans* and alternative formats, such as *IOB/CONLL*, yield identical results in terms of entity recognition, the *Prodigy-style spans* format aligns more closely with the unique characteristics of historical travelogue data and the annotation process. As a result, it offers a more intuitive and practical framework for this research, facilitating accurate and efficient entity extraction and analysis.

The performance of LLMs is automatically scored based on how closely their generated annotations align with the gold-standard annotations. Multiple evaluation criteria are applied to analyze the models' capabilities and limitations comprehensively. The evaluation process considers two primary aspects:

Span Identification: Assessing whether the model can accurately identify the exact span of an entity, regardless of the assigned type.

Entity Type Assignment: Evaluate whether the model correctly assigns the appropriate entity type, even if the span boundaries are not precise.

These metrics provide a balanced approach to scoring, addressing both the precision of entity boundaries and the correctness of type assignment, ensuring a thorough assessment of the models' annotation performance. The following five metrics are used to consider different categories of errors [54]:

Table 8.1: Error Types and Explanations

Error Type	**Explanation**
Correct (COR)	Both are the same
Incorrect (INC)	The output of a system and the golden annotation don't match
Partial (PAR)	System and the golden annotation are somewhat "similar" but not the same
Missing (MIS)	A golden annotation is not captured by a system
Spurious (SPU)	System produces a response which doesn't exist in the golden annotation

Furthermore, these five metrics are measured in four different ways [55]:

Table 8.2: Evaluation Schemas and Explanations

Evaluation Schema	**Explanation**
Strict	Exact boundary surface string match and entity type
Exact	Exact boundary match over the surface string, regardless of the type
Partial	Partial boundary match over the surface string, regardless of the type
Type	Some overlap between the system-tagged entity and the gold annotation is required

The performance of Large Language Models (LLM) in the task is measured using metrics that account for different scenarios of annotation accuracy: correct, incorrect, partial, missed, and spurious annotations. These metrics, as defined by the Message Understanding Conference (MUC), categorize and evaluate the results based on the following criteria [55]:

Table 8.3: Scenarios with Golden Standard, System Predictions, and Evaluation Schema Results

Scen.	**Golden Standard**		**System Prediction**		**Evaluation Schema**			
	Entity Type	**Surface String**	**Entity Type**	**Surface String**	**Type**	**Partial**	**Exact**	**Strict**
III	Action	departed			MIS	MIS	MIS	MIS
II			PER	He	SPU	SPU	SPU	SPU
V	PER	Mr. Stewart	PER	Stewart	COR	PAR	INC	INC
IV	PLACE	Argos	PER	Argos	INC	COR	COR	INC
I	PLACE	London	PLACE	London	COR	COR	COR	COR
I	PLACE	Germany	PLACE	Germany	COR	COR	COR	COR
VI	PER	John Stewart	PLACE	John Stewart, Esq.	INC	PAR	INC	INC

Evaluation Matrics

Precision, recall, and F1-score are calculated for each evaluation schema. To compute these metrics, two quantities need to be determined [54]:

$$\text{POSSIBLE (POS)} = \text{COR} + \text{INC} + \text{PAR} + \text{MIS} = \text{TP} + \text{FN}$$
$$\text{ACTUAL (ACT)} = \text{COR} + \text{INC} + \text{PAR} + \text{SPU} = \text{TP} + \text{FP}$$

Here, **TP** (True Positives) represents correct predictions, **FP** (False Positives) are incorrect predictions, and **FN** (False Negatives) are missed predictions.

Metrics Definition

Precision measures the percentage of correct named entities found by the NER system. In contrast, recall measures the percentage of named entities in the golden annotations retrieved by the NER system. These metrics are computed differently depending on whether an exact match or partial match scenario is required.

Exact Match (Strict and Exact)

For exact match scenarios, precision and recall are calculated as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{COR}}{\text{ACT}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{COR}}{\text{POS}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned}$$

Partial Match (Partial and Type)

For partial match scenarios, precision and recall incorporate partial matches (PAR) with a weighting factor of 0.5:

$$\begin{aligned} \text{Precision} &= \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{ACT}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{Recall} &= \frac{\text{COR} + 0.5 \times \text{PAR}}{\text{POS}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned}$$

The F1-score, which balances precision and recall, is calculated as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Bringing all components together, the final evaluation results are summarized in [Table 8.4].

Table 8.4: Evaluation Metrics

Measure	Type	Partial	Exact	Strict
Correct	3	3	3	2
Incorrect	2	0	2	3
Partial	0	2	0	0
Missed	1	1	1	1
Spurious	1	1	1	1
Precision	0.5	0.66	0.5	0.33
Recall	0.5	0.66	0.5	0.33
F1	0.5	0.66	0.5	0.33

These formulas ensure a comprehensive evaluation of the proposed annotation guidelines and LLM model's performance under both exact and partial match conditions.

8.2 Results

This study systematically evaluates the performance of the participating LLM models by comparing their outputs against a **gold-standard annotated dataset**. The primary objective is to assess the extent to which LLMs can approximate expert annotations through carefully designed prompts without direct optimization on the task. This evaluation offers insights into the capabilities and limitations of LLMs in historical data analysis, highlighting their potential to reduce dependence on extensive task-specific training data.

To ensure a fair and standardized comparison, a **structured prompting approach** is adopted for each model. Each prompt consists of three key components: **(i)** a detailed task description, **(ii)** two representative examples illustrating correct annotation strategies, and **(iii)** an explicit justification for each label assignment. This approach systematically assesses how well LLMs generalizes to the annotation task based solely on in-context learning.

Furthermore, a **Few-Shot Learning approach** is implemented on **GPT-4o** to evaluate its performance further [Figure 8.2]. This decision is motivated by two key constraints: **(i)** the OpenAI GPT API cost, which limits large-scale experimentation across multiple models, and **(ii)** the restricted context size of other models, which constrains their ability to incorporate various examples within a single prompt effectively. By leveraging a Few-Shot setup, this study aims to analyze whether providing additional context improves **GPT-4o's** ability to generalize and refine its entity annotation capabilities.

Table 8.5 summarizes the overall performance results, which presents a comparative analysis of all models in the Two-Shot setting. This evaluation provides a quantitative perspective on each model's ability to capture historical travel connections, recognize spatial relationships, and annotate textual entities with precision and recall. The subsequent sections further analyze the models'

relative strengths, weaknesses, and implications for their application in large-scale historical text annotation.

Table 8.5: LLMs F1-Scores Comparison (Two-Shot Learning)

LLM Models	Type	Partial	Strict	Exact
GPT-4o	63.73%	72.98%	62.50%	70.30%
GPT-4-turbo	55.19%	65.20%	54.30%	63.47%
GPT-3.5-turbo	31.23%	43.07%	30.80%	42.10%
LLaMA3:8B	03.21%	12.78%	0.00%	0.07%
Mistral:7B	02.46%	13.57%	0.00%	0.00%

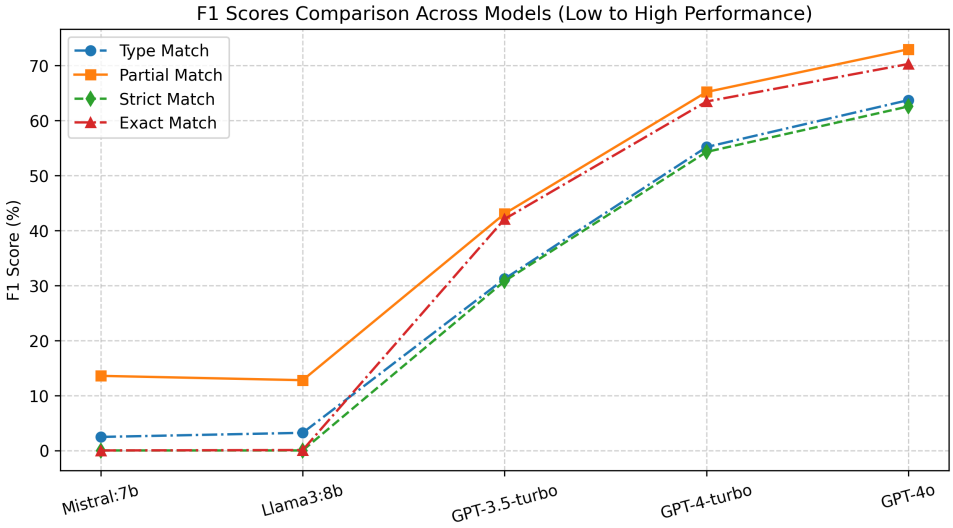


Figure 8.1: F1 Scores Comparison Across LLM Models.

8.2.1 Results Discussion

GPT-4o demonstrates the highest performance among all evaluated models, achieving strong results across multiple evaluation metrics. It attains an F1-score of **63.73%** in the **Type Metric**, with a precision of **72.88%** and a recall of **56.62%**, indicating a relatively high precision but a lower recall, which limits its ability to capture all relevant entities. The model performs significantly better in the **Partial Metric**, where its F1-score increases to **72.98%**, supported by a precision of **83.46%** and a recall of **64.84%**. This suggests that GPT-4o effectively identifies partial matches, improving its recognition of entity spans with some degree of flexibility. However, its performance declines in the **Strict Metric**, where the F1-score drops to **62.50%**, reflecting challenges in precisely identifying both entity boundaries and entity types. In the **Exact Metric**, GPT-4o achieves an F1-score of **70.30%**, demonstrating moderate accuracy in exact entity recognition. Overall, GPT-4o performs well in Two-Shot settings. While its high precision suggests strong entity identification capabilities,

the lower recall indicates that the model struggles to capture all relevant occurrences, which could impact its effectiveness in large-scale annotation tasks requiring comprehensive entity retrieval.

GPT-4-turbo emerges as the second-best performer among the evaluated models, though it demonstrates notable recall limitations. It achieves an F1-score of **55.19%** in the **Type Metric**, with a high precision of **76.00%** but a significantly lower recall of **43.33%**, indicating a tendency to correctly classify entities when detected but failing to capture a substantial portion of them. The model performs better in the **Partial Metric**, where its F1-score increases to **65.20%**, suggesting improved recognition of partial matches, with recall increasing to **51.19%**. However, in the **Strict Metric**, the F1-score declines to **54.30%**, highlighting inconsistencies in accurately identifying entity boundaries and types. The model demonstrates a relatively stronger performance in the **Exact Metric**, where it achieves an F1-score of **63.47%**, indicating moderate capability in recognizing entities with exact span matches. Overall, **GPT-4-turbo** exhibits a balanced but constrained performance, excelling in precision while struggling with recall. Its inability to consistently detect all relevant entities suggests that while it effectively minimizes false positives, it may omit a considerable number of valid entity mentions, limiting its reliability for exhaustive annotation tasks.

GPT-3.5-turbo exhibits significant weaknesses in recall and entity consistency, making it a less reliable choice for accurate entity recognition. In the **Type Metric**, the model achieves an F1-score of **31.23%**, with a precision of **67.58%** but a critically low recall of **20.30%**. This substantial gap between precision and recall indicates that while GPT-3.5-turbo is capable of correctly identifying entities when it does detect them, it fails to recognize a large portion of relevant instances. The **Partial Metric** shows a modest improvement, with the F1-score increasing to **43.07%**, suggesting slightly better performance in capturing partial entity matches. However, in the **Strict Metric**, the F1-score drops further to **30.80%**, reflecting poor accuracy in delineating entity boundaries and maintaining type consistency. The model demonstrates inconsistent recognition in the **Exact Metric**, where it attains an F1-score of **42.10%**, indicating that it struggles to produce reliable exact matches. Overall, **GPT-3.5-turbo** faces severe recall deficiencies, failing to detect a substantial proportion of the expected entities. While its precision remains moderate, its inability to comprehensively identify and classify entities suggests that it is not well-suited for tasks that require exhaustive entity retrieval and high recall rates.

LLaMA3 (Ollama version) exhibits abysmal performance, failing to generate meaningful annotations or effectively recognize named entities. The model achieves an F1-score of only **3.21%** in the **Type Metric**, with a precision of **4.99%** and an alarmingly low recall of **2.36%**, indicating that it not only misclassifies most detected entities but also fails to identify the vast majority of them. The **Partial Metric** shows a slight improvement, with an F1-score of **12.78%**, but both precision and recall remain critically low, suggesting a fundamental deficiency in recognizing entity spans even under relaxed criteria. In the **Strict Metric**, the F1-score drops to **0.00%**, demonstrating a complete inability to classify entities with the correct boundaries and types. The **Exact Metric** also

remains near zero, with an F1-score of only **0.07%**, highlighting the model's failure to generate precise entity annotations. Overall, **LLaMA3 is ineffective for NER tasks**, displaying near-zero performance across all evaluation metrics. Its inability to differentiate between entities and non-entities, combined with its extremely low recall, suggests that it is not viable for historical text annotation or any task requiring structured entity recognition.

Mistral (Ollama version) demonstrates slightly better performance than LLaMA3 in the *Partial Metric*, but it remains largely ineffective in NER tasks. The model achieves an F1-score of only **2.46%** in the **Type Metric**, with a precision of **6.08%** and a recall of just **1.54%**, indicating a significant failure in both entity detection and classification. While its performance improves slightly in the **Partial Metric**, reaching an F1-score of **13.57%**, this remains far below acceptable levels, and its precision-recall balance suggests that it struggles to recognize entity spans effectively. In the **Strict Metric**, the F1-score remains at **0.00%**, reflecting a complete inability to classify entity boundaries and types correctly. Similarly, the **Exact Metric** records an F1-score of **0.00%**, confirming that the model fails to provide precise entity annotations. Overall, **Mistral exhibits negligible improvements over LLaMA3 but remains unsuitable for NER tasks** in Two-Shot learning settings. Its low recall and lack of structured entity recognition suggest it is not a viable alternative for annotation tasks requiring accuracy and consistency in named entity classification.

In the Two-Shot setting, **GPT-4o** remains the most effective model, followed by **GPT-4-turbo**, while **GPT-3.5-turbo and quantized LLM models** struggle significantly. These findings underscore the importance of **model selection and contextual input examples** in achieving robust entity recognition performance in historical travelogue annotation.

8.2.2 Few-Shot Learning Performance

In the **Few-Shot** learning setting, GPT-4o achieves the highest overall performance among the evaluated models, demonstrating substantial improvements in entity recognition when provided with additional sample examples. The model attains an **F1-score of 81.27%** under the *Type* metric, with a **precision of 85.03%** and a **recall of 77.83%**, indicating a strong balance between accuracy and coverage in entity identification. A similar trend is observed across other evaluation metrics, with an **F1-score of 86.99%** under the *Partial* metric and **85.73%** under the *Exact* metric, suggesting a significant enhancement in the model's ability to correctly classify entity boundaries while minimizing both false positives and false negatives.

These results highlight the advantages of **Few-Shot learning**, as the model benefits from additional contextual examples, leading to improved generalization and annotation accuracy across different levels of entity recognition. The comparison between the **Few-Shot** and **Two-Shot** settings is visually represented in Figure 8.2, which illustrates the substantial performance gap between the two approaches.

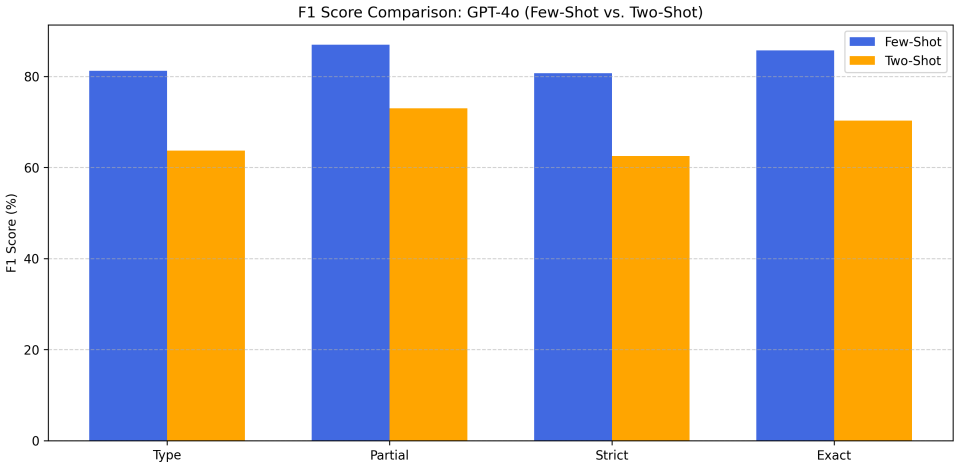


Figure 8.2: F1 Score Comparison: GPT-4o (Few-Shot vs. Two-Shot).

As shown in Figure 8.2, the **Few-Shot** learning approach consistently outperforms the **Two-Shot** setting across all metrics. The most significant improvements are observed in the *Partial* and *Exact* metrics, where the F1-score increases by approximately 14-19%. The results further validate the effectiveness of Few-Shot prompting in enhancing the model's ability to capture complex entity relationships, thereby improving annotation accuracy in historical travelogue analysis.

8.2.3 Itinerary Comparison

The comparison between the **gold annotation** and the **GPT-4o (Few-Shot)** annotated itinerary list for *BookID 001* (as referenced in Table 6.2) provides valuable insights into the model's ability to reconstruct historical travel sequences. The results in Table 8.6 demonstrate that **GPT-4o effectively captures key travel locations**, particularly in the initial stages of the itinerary, where it accurately identifies destinations such as **East Indies, England, India, Persia, Turkey, and London**. This alignment suggests that the model can recognize frequently occurring locations and reasonably follow historical travel patterns. However, notable deviations emerge in later itinerary entries, where the model misattributes locations, such as predicting **America instead of England** and **Paris instead of New York**. These inconsistencies indicate challenges in maintaining **long-range dependencies** and precise chronological ordering within complex travel narratives.

Additionally, **GPT-4o overgenerates locations** that do not appear in the gold annotations, introducing places such as **America**. This phenomenon likely results from **contextual interpolation**, where the model infers missing information rather than strictly adhering to the given data. While some of these locations may still be **historically plausible**, their absence from the gold standard suggests a deviation from **verified historical records**. Conversely, certain **gold-annotated locations are missing** or appear at different points in the model-generated itinerary, further underscoring

the model's difficulty in accurately structuring travel sequences.

Overall, while **GPT-4o (Few-Shot) demonstrates strong recall of frequently mentioned travel destinations**, its **inconsistent sequence alignment, omitted locations, and the introduction of inferred places** reveal limitations in itinerary reconstruction. These findings highlight the necessity of **human verification and post-processing** to refine and validate automated travel annotations, ensuring their historical accuracy and contextual reliability.

Table 8.6: Itinerary Comparison

Itinerary No.	Gold Annotation	GPT-4o (Few-Shot)
1	East Indies	East Indies
2	England	England
3	India	India
4	India	India
5	Persia	Persia
6	Turkey	Turkey
7	London	London
8	England	America
9	Scotland	England
10	Germany	England
11	Italy	Scotland
12	France	Germany
13	Paris	Italy
14	Paris	France
15	London	Paris
16	New York	Paris
17	Margate	England
18	London	London
19		America
20		New York
21		England
23		Margate
24		London

Finally, **GPT-4o (Few-Shot)** emerges as the most effective model, demonstrating robust precision, recall, and F1-scores across all evaluation metrics. It's Few-Shot learning capability enables it to effectively leverage minimal contextual examples, making it particularly well-suited for complex tasks requiring multiple reasoning steps. However, performance declines significantly for all models when transitioning to a Two-Shot setting, underscoring their sensitivity to the number of examples provided. **GPT-4-turbo** exhibits moderate performance but falls short compared to **GPT-4o**, particularly in terms of recall. In contrast, **GPT-3.5-turbo** and the *Ollama* models perform poorly, struggling with recall and maintaining entity consistency, which limits their applicability for Named Entity Recognition (NER) tasks under the specified guidelines and evaluation metrics. These findings highlight the critical importance of selecting models with strong generalization

capabilities and contextual understanding, particularly for NER tasks in domain-specific or structured annotation scenarios. The results underscore the need for models that can effectively balance precision and recall while adapting to the nuances of complex datasets.

8.2.4 ATO Results and Discussion

This study evaluates the performance of two different types of LLM models—**GPT-4o** and **LLaMA3:8B**—on the **ATO sample dataset** to assess their capability in structured annotation tasks. The results presented in Table 8.7 highlight the substantial performance differences between these models across four evaluation metrics: **Type, Partial, Strict, and Exact**.

Table 8.7: Comparison of LLM Performance on ATO Dataset (F1-Scores)

LLM Models	Data	Type	Partial	Strict	Exact
GPT-4o	Letter IV	56.67	56.39	50.00	52.78
	Letter V	54.88	54.27	51.22	52.44
LLaMA3:8b	Letter IV	02.15	07.17	0.00	0.00
	Letter V	04.44	04.44	02.22	02.22

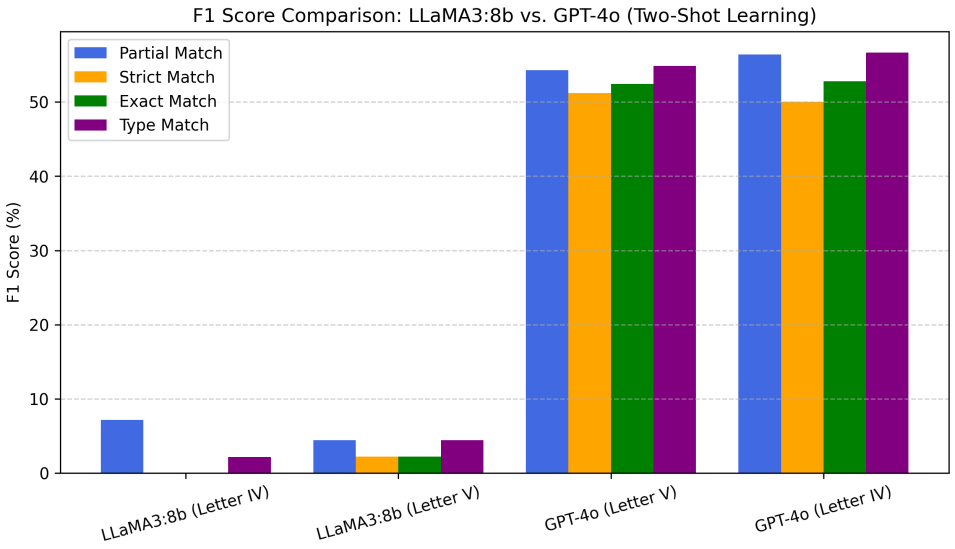


Figure 8.3: Performance Comparison of GPT-4o vs. LLaMA3:8B

Performance Comparison

Table 8.3 summarizes the overall performance results, which presents a comparative analysis of GPT-4o and LLaMA3:8B (quantized version) models in the Two-Shot setting.GPT-4o significantly outperforms LLaMA3:8B across all metrics. For **Letter IV**, **GPT-4o** achieves an F1-score of **56.67%**

under the **Type** metric, while **LLaMA3:8B** struggles at **2.15%**. A similar trend is observed for **Letter V**, with GPT-4o scoring **54.88%** and LLaMA3:8B only **4.44%**. The **Partial, Strict, and Exact metrics** follow the same pattern, with GPT-4o maintaining consistently higher scores, while LLaMA3:8B remains near zero, particularly in **Strict and Exact matching**, where it fails to provide meaningful annotations.

Model Suitability for Annotation Tasks

The vast performance gap between the models suggests that **GPT-4o is highly compatible with the annotation guidelines**, effectively recognizing entity relationships within the dataset. Conversely, **LLaMA3:8B struggles to generalize the annotation conventions, misclassifying named entities and failing to maintain structured annotations**, making it an unsuitable choice for historical text processing without additional fine-tuning.

Figure 8.3 visually illustrates these disparities, reinforcing the importance of selecting high-performance models for structured entity recognition tasks. The study's choice to evaluate **one advanced commercial model (GPT-4o) and one smaller open-source model (LLaMA3:8B)** highlights the significant gap between these architectures in handling annotation complexity.

These results confirm that **larger, high-context models like GPT-4o** are better suited for annotation tasks requiring complex entity recognition, while **smaller models like LLaMA3:8B struggle with accuracy and recall**. The study further demonstrates that while the developed **annotation guidelines are effective**, their compatibility with different datasets depends on the underlying model's ability to interpret entity relationships.

8.3 Error Analysis

The evaluation of quantized variants of large-scale LLM models reveals several limitations, particularly in the context of complex annotation tasks. These models are optimized for reduced computational requirements, allowing them to operate efficiently on limited GPU or CPU resources while maintaining acceptable performance levels. However, their constrained architecture results in difficulties when performing intricate reasoning tasks such as annotation based on stringent guidelines. Additionally, their limited context window inhibits the ability to retain and utilize annotations consistently across an entire book-length document.

The following subsections highlight key challenges identified in the performance of quantized models, along with specific examples illustrating their impact on annotation accuracy.

8.3.1 Limitations in Model Capabilities for Complex Annotation Tasks

The inability of these models to fully adhere to annotation guidelines results in frequent misclassifications and structural inconsistencies. As these models are primarily trained for tasks such

as conversational AI, code generation, and basic NER for commonly recognized entities, their performance in domain-specific annotation is often inadequate.

For example, in the following instance, the model incorrectly placed an <Action> tag, initiated a <PER> tag without properly closing it and subsequently misclassified a person entity as a place entity [Table 8.8]:

Table 8.8: Comparison of Original vs. Generated Data

Original Data	**Generated Data (LLaMA3:8B)**
"They are presented to the public as they were at intervals elicited from his excellent departed friend and relative, John Stewart, Esq. was born in London in the year 1749." [18]	"They are presented to the public as they were at intervals elicited from **<Action SpatialRelation=FROM>**his excellent departed friend and relative</Action>, **<PER id=0><PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False>**John Stewart, Esq.</PLACE> was born in <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False>London</PLACE> in the year 1749." [18]

8.3.2 Inability to Maintain the Specified Output Structure

The models, particularly those with limited context size, struggle to maintain the required output structure when processing multi-page inputs. Once the model reaches its context window limit, it begins discarding earlier context, leading to annotation inconsistencies. In contrast, larger models such as GPT-4o effectively maintain output integrity over longer sequences as shown in the following example.

GPT-4o Generated Data:

"*<PER id=0>He</PER> made the tour of <PLACE TemporalSpan=HABITUAL, IsPartOfJourney=TRUE, Itinerary_ID=14>Scotland</PLACE>, <PLACE TemporalSpan=HABITUAL, IsPartOfJourney=TRUE, Itinerary_ID=15> Germany</PLACE>, <PLACE TemporalSpan=HABITUAL, IsPartOfJourney=TRUE, Itinerary_ID=16>Italy</PLACE>, and <PLACE TemporalSpan=HABITUAL, IsPartOfJourney=TRUE, Itinerary_ID=17>France</PLACE>, on foot, and ultimately <Action SpatialRelation=TO>settled in</Action> <PLACE TemporalSpan=HABITUAL, IsPartOfJourney=TRUE, Itinerary_ID=18>Paris </PLACE>, where <PER id=0>he</PER> quickly gained the friendship of the leaders of all parties, having their confidence;—a proof how greatly <PER id=0>his</PER> moral character was esteemed.*" [18]

LLaMA3:8B Generated Data:

"He made the tour of Scotland, Germany, Italy, and France, on foot, and ultimately settled in Paris, where he quickly gained the friendship of the leaders of all parties, having their confidence;—a proof how greatly his moral character was esteemed." [18]

8.3.3 Failure to Detect Historical Place Entities

Due to their compressed size and limited knowledge base, these models fail to recognize historical place names that are uncommon or no longer in use, leading to annotation inaccuracies. While GPT-4o successfully identifies these entities, LLaMA3:8B frequently misclassifies them as person entities [Table 8.9].

Table 8.9: Comparison of LLMs Generated Data

Generated Data (GPT-4o)	Generated Data (LLaMA3:8B)
"In this manner stood the cities of <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False>Argos</PLACE>, <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False>Sicyon</PLACE>, <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False>Corinth</PLACE>, ... " [43]	"In this manner stood the cities of **<PER id=5>**Argos</PER>, <PER id=6>Sicyon </PER> , <PER id=7>Corinth</PER>, ... " [43]

8.3.4 Omissions of Action Tags

Another common issue observed in quantized models is the omission of required action tags, even when explicit keywords were provided in the instructions. For instance, in the following example, the model failed to recognize and annotate the <Action> tag:

LLaMA3:8B Generated Data: *"On the 23rd of September, I **departed** from <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=True>Bucharest</PLACE>."* [20]

8.3.5 Introduction of Unauthorized Tags and Invalid Spatial Relation Values

Certain models also generate unauthorized tags and incorrect spatial relation values, reducing annotation consistency. For example, in the following case, the model introduced an unauthorized PLACETYPE tag and an invalid spatial relation value (SpatialRelation=WITH):

LLaMA3:8B Generated Data:

- *"The war with **<PLACETYPE EthnicGroup>**Austria**</PLACETYPE>** not permitting me to*

pass through <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=False> Transylvania</PLACE>." [20]

- "*As we had coffee prepared for ourselves, Mukni, who was fond of it, visited us in our tent, <Action SpatialRelation=**WITH**>and partook of it with his son <PER id=7, ValidConnection=True, TravelConnection=False>Yussuf</PER>, a mulatto boy of about eight years of age.</Action>*" [20]

8.3.6 Modifications or Omissions of Original Text

Despite explicit instructions to preserve the original text, certain models tend to modify or omit parts of the input, leading to potential data integrity issues. In the following example, the model removed the phrase *"THE LIFE OF"* from the original text, resulting in incomplete output:

Original Data: "***THE LIFE OF** JOHN STEWART, Esq. When a man of genius, remarkable for talent or noted for eccentricity, leaves this world forever and ever, the public mind immediately...*" [18]

Generated Data (LLaMA3:8B): "*<PER id=0>**John Stewart**, Esq.</PER> When a man of genius, remarkable for talent or noted for eccentricity, leaves this world forever and ever, <PER id=1, ValidConnection=False>the public mind</PER> immediately...*" [18]

8.3.7 Errors in Numbering Person Identifiers

The models also exhibit inconsistencies in assigning unique person identifiers. In some instances, the same individual is assigned different IDs within a single passage, leading to inaccurate annotations.

LLaMA3:8B Generated Data: "*It is true that <PER id=11>M. Stewart</PER> 's new system of things savored of a little tincture of romance. However, <PER id=0>M. Stewart</PER> met with a few who were willing to yield him that credence...*" [18]

The error analysis highlights fundamental challenges in applying quantized LLMs for complex annotation tasks. Issues such as limited recall, annotation structure inconsistencies, omission of critical tags, and incorrect entity classifications indicate that these models require further optimization. Future work should focus on fine-tuning models on domain-specific datasets and integrating hybrid AI-human annotation workflows to enhance the robustness and accuracy of automated historical text analysis.

8.4 Discussion

The results demonstrate the compatibility of the proposed annotation guidelines for **valid itinerary extraction** in historical travelogues, confirming their effectiveness in structuring and interpreting travel sequences within historical narratives. This validation underscores the utility of the proposed

framework for analyzing historical travelogues, offering a fundamental methodology for extracting meaningful insights from complex and nuanced narratives. While the guidelines yield promising results, establishing a **valid connection between a character and a specific location** requires both **the character and the location to be explicitly referenced in the text**. Implicit references or inferred locations pose a challenge for LLM models, necessitating more sophisticated reasoning capabilities or human intervention to ensure annotation accuracy.

Furthermore, the utilization of LLM models highlights their capability in **annotation tasks through in-context learning**, allowing them to perform structured entity recognition without extensive fine-tuning. The performance evaluation of different models reveals a stark contrast between **high-capacity models and quantized alternatives**. **GPT-4o outperforms all other models, achieving high F1-scores across Type, Partial, Strict, and Exact metrics**, while **quantized open-source models such as LLaMA3:8B and Mistral:7B struggle significantly**, demonstrating **poor recall and annotation inconsistencies** across datasets.

The study also underscores the adaptability of **annotation guidelines across different datasets**, where **GPT-4o maintains strong results**, albeit with slight variations due to differences in text structure, entity density, and language. In contrast, **quantized models exhibit consistently poor performance**, indicating that **smaller models require domain-specific fine-tuning** to align with structured annotation frameworks.

Finally, this research contributes to future advancements in **historical travelogue itinerary extraction** by demonstrating how LLM models can be leveraged through **in-context learning for domain-specific tasks**. The results emphasize the need for **human supervision in the annotation loop** to handle ambiguous cases and **ensure data integrity**. Additionally, **quantized versions of LLM models require further fine-tuning and optimization** to improve their applicability in **automated historical text annotation**, making them more viable for large-scale deployment in domain-specific contexts.

Chapter 9

Conclusion

This research presents a comprehensive methodology for mapping and analyzing travel connections between *person and place entities* in English historical travelogues [More details on this can be found in Chapter 5]. By addressing the limitations of traditional Named Entity Recognition (NER) models, which often treat *person and place entities as independent units*, this study introduces a novel framework that classifies *places based on their significance within a journey*, thereby capturing *spatial, temporal, and contextual relationships* in historical narratives. The development of **precise annotation guidelines** and the compilation of a *specialized corpus* represent an **initial effort toward** the *integrated modeling of person and place entities*, providing historians and researchers with a more structured approach to identifying, verifying, and analyzing *historical travel routes*.

The sample examples presented in Chapter 5 demonstrate that **combining spatial, temporal, and linguistic contextual cues** allows for a more holistic understanding of historical travelogues. By systematically associating *travelers with locations*, researchers can distinguish *key journey stops from incidental place mentions, refining historical itinerary reconstruction*. This refined approach significantly enhances *historical travel route extraction*, equipping historians and digital humanities researchers with *new computational tools to validate and expand existing travel narratives*.

A key component of this research is the **application of LLM models and automated annotation processes** to extract and classify travel-related entities. By employing *structured prompt engineering, model customization, and post-processing validation*, the study demonstrates how OpenAI **GPT-based models, quantized open-source LLaMA, and Mistral models** can automate travel annotation with varying levels of accuracy. The integration of *'Modelfiles'*, structured inference pipelines, and **Nervaluate**-based evaluation highlights the *efficacy and limitations* of LLM-based approaches for *complex historical text annotation*. While OpenAI *GPT* models exhibit superior instruction-following capabilities, quantized open-source *LLaMA* and *Mistral* models provide cost-effective, locally deployable alternatives, albeit with reduced contextual reasoning abilities. The *scalability of automated annotation pipelines*, combined with resource optimization and error-handling mechanisms,

demonstrates the feasibility of large-scale historical text analysis using modern LLM architectures.

The **proposed annotation guidelines** significantly improve annotation consistency and interpretability, establishing a robust framework for **travel entity classification**. The guidelines standardize annotation procedures across diverse travel narratives by structuring *spatial relations, travel connections, and journey segmentation rules*. This structured approach enhances data interoperability, facilitating comparative studies across historical sources. Furthermore, the **integration of Nervaluate-based evaluation metrics** validates annotation accuracy, demonstrating the effectiveness of the proposed methodology in *capturing historical movement patterns*.

In conclusion, this research lays the groundwork for *future advancements in historical text annotation and travelogue analysis*. While the study presents a *scalable, automated framework*, several challenges remain, including *instruction-following inconsistencies in quantized LLM models*, *historical entity recognition gaps*, and *computational constraints in long-context processing*.

9.1 Limitations

Two significant limitations exist in the conceptualization of the main task. Firstly, to establish a connection between a character and a specific location, it is necessary for both *the character and the location to be explicitly referenced in the text*. This simplification aids in annotation and modeling but falls short when considering the broader context of grounding any character at any moment within the narrative. Another limitation is the model's potential *inability to distinguish between planned journeys and actual journeys* within complex sentence structures lacking clear indicators. This limitation can adversely impact the accuracy of itinerary classification.

9.2 Future Work

Future work should focus on **fine-tuning quantized LLM models or integrating hybrid annotation strategies for domain-specific tasks** with *historical travel databases* to enhance rare entity recognition and reduce annotation inconsistencies. Extending *Nervaluate-based evaluation metrics* will also enable more granular *domain-specific annotation consistency checks*. The **integration of hybrid annotation strategies**, combining rule-based systems with LLMs, may offer improvements in precision and reliability, addressing current limitations in *historical entity tagging and spatial relation classification*. Furthermore, **optimizing inference efficiency** through context-aware segmentation and distributed processing architectures will help reduce computational costs while maintaining high-quality annotations. Moreover, the itineraries can be **visualized on maps** by leveraging the tool functions of the LLM models. This study tested this feature in the preliminary stage with a domain-specialized GPT-4 model, referred to as the 'Travelogue Annotator' version-1 [52]. This approach demonstrates the potential for integrating advanced visualization tools with annotation workflows, enabling more intuitive and accessible representations of historical travel narratives.

Future research can further refine *historical travel analysis* by building on these findings, contributing to digital humanities and computational linguistics.

Bibliography

References

[2] A. Hamdi, E. L. Pontes, E. Boros, *et al.*, "A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers," in *In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, 2328–2334. [Online]. Available: https://doi.org/10.5281/zenodo.4573313.

[3] S. Soni, A. Sihra, E. Evans, M. Wilkens, and D. Bamman, "Grounding characters and places in narrative text," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, 2023, pp. 11 723–11 736. [Online]. Available: https://aclanthology.org/2023.acl-long.655.

[4] S. Balck, H. Beyer-Thoma, F. Ingo, and A. Ananieva, "Interlinking text and data with semantic annotation and ontology design patterns to analyse historical travelogues," in *Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices*, Online: Digital Humanities Quarterly 17 (3), 2023. [Online]. Available: https://www.digitalhumanities.org/dhq/vol/17/3/000726/000726.html.

[5] J. Beck, "Quality aspects of annotated data: A research synthesis," *AStA Wirtschafts- und Sozialstatistisches Archiv*, vol. 17, p. 1, Nov. 2023. DOI: 10.1007/s11943-023-00332-y.

[6] J.-P. Rubiés, "Travellers and cosmographers," in Routledge, 2023, ch. [1], [7–10]. [Online]. Available: https://www.perlego.com/book/4174844.

[7] K. Hosseini, T. Kober, J. Krapac, R. Vollgraf, W. Cheng, and A. P. Ramallo, *Retrieve, annotate, evaluate, repeat: Leveraging multimodal llms for large-scale product retrieval evaluation*, 2024. arXiv: 2409.11860 [cs.IR]. [Online]. Available: https://arxiv.org/abs/2409.11860.

[8] W. Hsieh, Z. Bi, C. Jiang, *et al.*, *A comprehensive guide to explainable ai: From classical models to llms*, 2024. arXiv: 2412.00800 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2412.00800.

[10] A. Ananieva, S. Balck, and J. Möhrke, "The study of historical travelogues from a digital humanities perspective: Experiences and new approaches," *Comparative Southeast European Studies*, vol. 72, no. 3, pp. 370–385, 2024. [Online]. Available: https://doi.org/10.1515/soeu-2024-0023.

[11] I. Mani, J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner, "Spatialml: Annotation scheme, corpora, and tools," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco: European Language Resources Association (ELRA), 2008. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2008/pdf/106_paper.pdf.

[12] P. Silvano, A. Leal, F. Silva, I. Cantante, F. Oliveira, and A. M. Jorge, "Developing a multilayer semantic annotation scheme based on ISO standards for the visualization of a newswire corpus," in *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Groningen, The Netherlands (online): Association for Computational Linguistics, Jun. 2021, pp. 1–13. [Online]. Available: https://aclanthology.org/2021.isa-1.1.

[13] A. A. S. B. J. Möhrke, "Developing geographically oriented nlp approaches to sixteenth–century historical documents: Digging into early colonial mexico," Online: Digital Humanities Quarterly 14 (4), 2020. [Online]. Available: http://www.digitalhumanities.org/dhq/vol/14/4/000490/000490.html#.

[14] M. C. Ardanuy, "Entity-centric text mining for historical documents," Ph.D. dissertation, Georg-August-University Goettingen, 2017.

[15] Y. Luan, D. Wadden, L. He, A. Shah, M. Ostendorf, and H. Hajishirzi, "A general framework for information extraction using dynamic span graphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3036–3046. DOI: 10.18653/v1/N19-1308. [Online]. Available: https://aclanthology.org/N19-1308/.

[16] N. Mostafazadeh, A. Grealish, N. Chambers, J. Allen, and L. Vanderwende, "Caters: Causal and temporal relation scheme for semantic annotation of event structures," in *Proceedings of the Fourth Workshop on Events*, San Diego, California: Association for Computational Linguistics, 2016, 51–61. [Online]. Available: https://aclanthology.org/W16-1007.

[17] N. Reiter and A. Frank, "Identifying generic noun phrases," in *In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Uppsala, Sweden: Association for Computational Linguistics, 2010, 40–49. [Online]. Available: https://aclanthology.org/P10-1005.

[22] N. Ide, J. Pustejovsky, H. Bunt, and A. Zaenen, "Handbook of linguistic annotation," in Springer Dordrecht, 2017, ch. [1-2], [1–3, 21–42]. [Online]. Available: https://doi.org/10.1007/978-94-024-0881-2.

[23] LDC, "Ace (automatic content extraction) english annotation guidelines for entities.," *Linguistic Data Consortium*, vol. 5.6.6, 2005. [Online]. Available: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf.

[24] D. Bamman, S. Popat, and S. Shen, "An annotated dataset of literary entities," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2138–2144. DOI: 10.18653/v1/N19-1220. [Online]. Available: https://aclanthology.org/N19-1220.

[25] T. Aarsen, *Spanmarker for named entity recognition*, Available at https://github.com/tomaarsen/SpanMarkerNER, 2023.

[26] H. Face, *Integrating hugging face and spacy: Combining models for nlp tasks*, Available at https://huggingface.co/blog/spacy, 2023.

[27] W. X. Zhao, K. Zhou, J. Li, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023. [Online]. Available: https://arxiv.org/abs/2303.18223.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[29] T. Eloundou, S. Manning, P. Mishkin, and D. Rock, "Gpts are gpts: An early look at the labor market impact potential of large language models," *arXiv preprint arXiv:2303.10130*, 2023.

[30] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[31] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[32] L. Ouyang, J. Wu, X. Jiang, *et al.*, *Training language models to follow instructions with human feedback*, 2022. arXiv: 2203.02155 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2203.02155.

[33] A. Q. Jiang, A. Sablayrolles, A. Mensch, *et al.*, *Mistral 7b*, 2023. arXiv: 2310.06825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.06825.

[34] I. John, "The art of asking chatgpt for high-quality answers: A complete guide to prompt engineering techniques," in Nzunda Technologies Limited, 2023, ch. [1, 5], [6–7, 14–15].

[35] D. Powers and Ailab, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation," *J. Mach. Learn. Technol*, vol. 2, pp. 2229–3981, Jan. 2011. DOI: 10.9735/2229-3981.

[36] T. A. Nisha, *Guidelines for corpus preparation and annotation of historical travelogues: A path to valid itinerary discovery*, Available at: https://gitlab.gwdg.de/tarjia.alamnisha/researchorientedproject.git, 2024.

[37] M. Recasens, E. Hovy, and M. A. Martí, "A typology of near-identity relations for coreference (nident).," in *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*, Valletta, Malta: European Language Resources Association (ELRA), 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/160_Paper.pdf.

[49] H. Zeghidi and L. Moncla, *Evaluating named entity recognition using few-shot prompting with large language models*, 2024. arXiv: 2408.15796 [cs.IR]. [Online]. Available: https://arxiv.org/abs/2408.15796.

[50] C. Ling, X. Zhao, J. Lu, *et al.*, *Domain specialization as the key to make large language models disruptive: A comprehensive survey*, 2024. arXiv: 2305.18703 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2305.18703.

[51] L. Cheng, F. Pannach, S. van Dijk, F. Pianzola, B. Ramakers, and E. van Rein, "The academic tourism ontology (ato): Fostering linked open data and information sharing in cultural heritage and literary studies," [Manuscript submitted for publication], 2025.

[52] T. A. Nisha, *Travelogue annotator*, OpenAI platform, 2024. [Online]. Available: https://chatgpt.com/g/g-M46uUIBD6-travelogue-annotator.

[53] T. A. Nisha, F. Pannach, and J. Wettlaufer, "Itinerare erkennen in Reiseberichten. Auszeichnung von Orts- und Personennamen zur Etablierung von Itineraren in Reiseberichten des 19. Jahrhunderts," in *DHd 2024 – Passau: Workshop, Generative KI, LLMs und GPT bei digitalen Editionen*, 2024. [Online]. Available: https://dhd2024.dig-hum.de/w8-generative-ki-llms-und-gpt-bei-digitalen-editionen.

[54] D. Batista and M. A. Upson, *Nervaluate (version 0.2.0)*, version 0.2.0, Computer software, 2020. [Online]. Available: https://github.com/mantisnlp/nervaluate.

[55] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, "SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)," in *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Atlanta, Georgia, USA: Association for Computational Linguistics, Jun. 2013, pp. 341–350. [Online]. Available: https://aclanthology.org/S13-2056/.

[56] P. J. Phillips, P. J. Phillips, C. A. Hahn, *et al.*, "Four principles of explainable artificial intelligence," 2021. [Online]. Available: https://doi.org/10.6028/NIST.IR.8312.

[57] E. Kearns, "Annotating and quantifying narrative time disruptions in modernist and hypertext fiction," in *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, Online: Association for Computational Linguistics, 2020, 72–77. [Online]. Available: https://aclanthology.org/2020.nuse-1.9.

[58] F. Aladağ, "The potential of gpt in ottoman studies: Computational analysis of evliya çelebi's travelogue with nlp and text mining and digital edition with tei," Preprint version, 2023. [Online]. Available: https://doi.org/10.13140/RG.2.2.10335.43683.

Travelogues

[1] R. Chandler, "Travels in asia minor , and greece: Or an account of a tour made at the expense of the society of dilettanti," in 3rd. Joseph Booker and R. Priestley, 1817, vol. 1, ch. [1-2], [1–8].

[18] W. T. Brande, "The life and adventures of the celebrated walking stewart: Including his travels in the east indies, turkey, germany, & america. by a relative. with a portrait," in E. Wheatley, English and Foreign Bookseller, 1822, [3–12].

[19] W. Macmichael, "Journey from moscow to constantinople, in the years 1817, 1818," in London, J. Murray, 1819, ch. [4], [187–207].

[20] J. M. Tancoigne, "A narrative of a journey into persia and residence at teheran, containing a descriptive itinerary from constantinople to the persian capital," in William Wright, 1820, ch. [34], [386–394].

[21] M. D. R. Richardson, "Travels along the mediterranean, and parts adjacent; in company with the earl of belmore, during the years 1816-17-18: Extending as far as the second cataract of the nile, jerusalem, damascus, balbec," in T. Cadell, 1822, vol. 2, ch. [19], [238–258].

[38] F. Beaufort, "Karamania: A brief description of the south coast of asia minor and the remains of antiquity.," in R. Hunter, Successor to Mr. Johnson, 1817, ch. [10], [194–211].

[39] C. de Forbin, "Travels in egypt, being a continuation of the travels in the holy land in 1817-18," in Sir Richard Phillips & Co., 1819, vol. 2, [1–14].

[40] H. A. Barker and J. Burford, "Description of the view of athens, and surrounding country," in Jas. Adlard and Sons, 1818, [3–12].

[41] C. G. F. Lyon, "A narrative of travels in northern africa, in the years 1818, 19, and 20," in John Murray, 1821, ch. [2], [38–75].

[42] J. M. A. Scholz, "Travels in the countries between alexandria and paraetonium, the lybian desert, siwa, egypt, palestine, and syria, in 1821," in Sir Richard Phillips & Co., 1822, [1–11].

[43] L. E. D. Clarke, "Travels in various countries of europe, asia and africa," in 4th ed. T. Cadell and W. Davies, 1817, vol. 3, ch. [4], [199–213].

[44] H. Light, "Travels in egypt, nubia, holy land, mount libanon, and cyprus, in the year 1814," in Rodwell and Martin, 1818, ch. [4], [101–118].

[45] H. A. S. Dearborn, "Memoir on the commerce and navigation of the black sea, and the trade and maritime geography of turkey and egypt," in Wells and Lilly, 1819, vol. 2, [272–294].

[46] J. H. Clark, "The military costume of turkey. illustrated by a series of engravings from drawings made on the spot: Dedicated by permission to his excellency the minister of the ottoman porte to his britannic majesty," in T. M'LEAN, 1819, [25–41], Contributors include Antonaki Ramadani, King George III, John Heaviside Clark, and Thomas Wageman. Author not mentioned.

[47] L. W. Heude, "A voyage up the persian gulf, and a journey overland from india to england, in 1817," in Strahan and Spottiswoode, 1819, ch. [4], [52–69].

[48] T. A. Nisha, *Travelogue corpus*, Available at: `https://gitlab.gwdg.de/tarjia.alamnisha/mastersthesis.git`, 2025.

Annotation Guidelines

Tarjia Alam Nisha

03.03.2025

Task Objective

The goal is to determine and describe the type of connection or relationship that exists between two specific characters or a specific character (Traveler) and a designated place/location within a given excerpt. This entails identifying the nature of their association and providing an explanation or description of it. (Which will help us to find valid itineraries of a traveler or group of travelers.)

Example:

> *John Stewart, Esq. was born in London in the year 1749. [18]*
>
> *I (WILLIAM MACMICHAEL) shall always reflect with great satisfaction on the adventures of a journey during which I enjoyed the pleasure of the society of my friend Mr. Legh. [19]*

What is a Character?

A character refers to a human or a person entity within a narrative that possesses the capability to engage in monologues or dialogues. An automated entity tagger is employed to identify characters, with entities categorized as `PER` (person) according to the guidelines specified in a particular research paper, being regarded as potential characters.

Example:

> *Mr. Stewart displayed the most fanciful and yet pleasing taste. [18]*

What is a Place?

A place refers to a defined spatial area that can be physically demarcated. It encompasses locations such as rooms, cities, or specific geographical features. Similar to character identification, candidates for place entities are also recognized automatically using an entity tagger. In this case, entities falling under the categories of `LOC` (location), `FAC` (facility), or `GPE` (geopolitical entity), as outlined in a specific research paper, are regarded as potential places.

Example:

> *Mr. Stewart returned to England (GPE). [18]*

Task Details and Setup

The excerpts will be provided from the source of English travelogues of the Ottoman Empire. Within each excerpt, there will be a highlighted section representing a character or characters, as well as a separate highlighted portion representing a place. The identification of these character

and place spans will be achieved through an automated system, eliminating the need for manual intervention in the process. If an annotator finds a place that was not correctly recognized by the tagger, then the annotator will provide the label themselves.

Seven annotations are to be provided by every annotator for any excerpt:

Annotation	Type (Domain)	Possible values
Valid relation [3]	Binary (PER-LOC)	• TRUE • FALSE
Spatial relation [3]	Categorical (PER-LOC)	• IN • NEAR • THRU (Through) • TO • FROM • NO REL
Temporal span [3]	Binary (PER-LOC)	• PUNCTUAL • HABITUAL
IsPartOfJourney	Binary (PER-LOC)	• TRUE • FALSE
Valid connection	Binary (PER-LOC)	• TRUE • FALSE
Travelling relation/ Travel relationship / Travel connection	Categorical (PER-PER)	• IN GROUP • JOIN GROUP • EXIT GROUP • NO REL

Example Annotation

Consider the following sentence:

”<PER id=0> Alex </PER> and <PER id=1, ValidConnection=True, TravelConnection=IN GROUP > Maya </PER> <Action SpatialRelation=FROM > started</Action> their adventurous journey <Action SpatialRelation=FROM>from</Action> the mountainside village called <PLACE TemporalSpan=PUNCTUAL, IsPartOfJourney=TRUE, Itinerary_ID=1> Hallstatt </PLACE>.”

For annotation purposes, we have assigned `id=0` inside the `<PER>` tag to the primary character, Alex (the central figure or protagonist in the narrative whose journey is of primary interest), and `id=1` to the other character, Maya. We have labeled the keywords that denote spatial relationships as `<Action>` and assigned an Itinerary_ID number to the `<PLACE>` where `IsPartOfJourney = TRUE`, in a way that represents the character's actual movement. Here, the annotation description for the characters Alex, Maya, and the place Hallstatt is as follows:

- **Valid Relation:** TRUE (Both characters and the place are correctly tagged.)
- **Spatial Relation:** FROM (Alex and Maya started the journey 'From' Hallstatt, indicating their departure from the place.)
- **Temporal Span:** PUNCTUAL (The journey is presented as a single, specific event.)
- **IsPartOfJourney:** TRUE (The narrative indicates the start of a journey, and they started from 'Hallstatt' together.)
- **Valid Connection:** TRUE (The narrative indicates a valid connection between the primary and the other characters.)
- **Travel Connection:** IN GROUP (The narrative describes Alex and Maya traveling together, which indicates a valid travel connection between the characters.)

This annotation setup ensures that the primary character's movement is accurately represented, with the spatial relationship clearly indicated through the labeled action and the place assigned an Itinerary_ID. The place 'Hallstatt' is part of the journey (`IsPartOfJourney = TRUE`) and is appropriately annotated to reflect its significance in the itinerary.

Annotation Descriptions

1. Valid relation (PER–LOC)

This is a binary annotation that indicates whether the character and the place are correctly identified by the entity tagger. If the character or the place are not correctly identified, the annotation should be `FALSE`.

2. Spatial relation (PER–LOC)

This is a categorical annotation that indicates the spatial relationship between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative; identified as `[id = 0]`) and the place. The possible values are:

- `IN`: The character is inside the place. Indicated using phrases such as "at," "arrived," "on," "into," "visited," "landed at," "attended in," "received," "remained at," "landed on," etc.

- NEAR: The character is near the place and must have a clear indication of it, using phrases such as "near," "approached near to," "encamped near," "bark opposite," "in the vicinity of," "nearby," etc.
- THRU: The character is passing through the place. Indicated using phrases such as "passing by," "passing through," "wafted by," "sailed by," "passed," "crossed," "across," "through," "over," etc.
- TO: The character is going *towards* the place, with an indication of arrival at the destination such as "landing," "entering to," "returned to," "came to," "brought to," "went to," etc.
- FROM: The character is coming from the place or started the journey from the place, with an indication of being at that location. Indicated using phrases such as "sailing from," "out of," "left," "departure," "travelled from," "returning from," "leaving," "escaped from," etc.
- NO REL: There is no spatial relationship between the primary character [id = 0] and the place.

3. Temporal span (PER–LOC)

This is a binary annotation that indicates whether the primary character's interaction with the place is punctual or habitual. If the primary character's interaction with the place is a one-time event or is mentioned only once in the context, the annotation should be designated as PUNCTUAL. Conversely, if the primary character's interaction with the place is repeated, ongoing, or involves a birth connection, the annotation should be HABITUAL.

4. IsPartOfJourney (PER–LOC)

This binary annotation determines whether the referenced location, associated with the primary character ([id = 0]), corresponds to a segment of that individual's own part of the journey or not. If the Spatial Relation is IN, FROM, TO, NEAR, or THRU, then the primary character's interaction with the place is part of the journey, and the annotation should be TRUE.

If IsPartOfJourney = TRUE, and the character is moving from A to B via X/Y, stays for some time or a few days in Z, and finally reaches the destination B, follow these steps to construct the itinerary representing the character's actual movement:

- **Starting Point (FROM):** If the Spatial Relation is FROM, which indicates the departure or start of the journey (such as A), then this place comes first in the itinerary.
- **Intermediate Locations (THRU or NEAR):** If the Spatial Relation is THRU or NEAR (such as X/Y), then these places come in the middle of the itinerary.
- **Stops or Stays (IN):** If the Spatial Relation is IN, which indicates being at a place for a few days or having stopped there for some time (such as Z), it should be placed accordingly in

the itinerary, reflecting the sequence of the journey.

- **Destination (TO):** If the Spatial Relation is `TO`, which indicates the destination place (such as B), then it comes at the end of the itinerary.

Therefore, the construction of the itinerary should be as follows:

$$A \rightarrow X/Y \rightarrow Z \rightarrow B$$

On the other hand, if the Spatial Relation is `NO REL` or the passage indicates a journey that transpired in an earlier period, distinct from the timeframe governing the current narrative, then the primary character's mentioned place is not part of the journey, and the annotation should be `FALSE`.

Furthermore, if the narrative matches any of the following conditions, then the primary character's mentioned place is not part of the journey, and the annotation should be `FALSE`:

Keywords and Phrases Indicating Future Intentions

- Verbs expressing intention or decision: *plan*, *intend*, *decide*, *prepare*, *arrange*, *schedule*, *aim*, *hope*, *expect*.
- Modal verbs suggesting future actions: *will*, *shall*, *would*, *should*, *could*.
- Phrases indicating consideration or discussion of travel: *thinking about traveling to*, *going to*, *drawing on*, *obliged to*, *discussed visiting*, *contemplating a journey to*, *changed my plan*, *decided on continuing my journey by land*, *set our sights on*.

Expressions of Preparation

- Logistical arrangements: Mention of booking passages, packing belongings, securing travel documents, or arranging accommodations.
- Financial preparations: Discussing budgets for the journey, exchanging currency, or financial planning related to travel expenses.
- Gathering information: Seeking maps, guides, or advice about destinations and routes or distance from the destination (e.g., fifty leagues from Constantinople).

Time References

- Future time indicators: *next week*, *soon*, *upcoming*, *in the coming months*, *eventually*, *afterward*.
- Conditional statements about future travel: *If the weather permits, we will...*, *Should the opportunity arise, I plan to...*.

Dialogues or Correspondence

- Conversations about travel plans: Characters discussing with others about where and when they intend to go or asking for a permit to visit the place.
- Letters or messages outlining intentions: Written plans or invitations related to future or past journeys.

Emotional Expressions

- Feelings about upcoming travel: Expressions of excitement, anxiety, or anticipation regarding a planned journey.
- Motivations for travel: Desires to explore, the need to relocate, or obligations requiring travel.

5. Valid connection (PER–PER)

This is a binary annotation that indicates whether there is a relationship between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and other characters or not. If the characters have no relationship, the annotation should be `FALSE`.

6. Travel connection (PER–PER)

This is a categorical annotation that indicates the travel relationship between the primary character (whose journey is of primary interest in the story or the central figure or protagonist in a narrative) and other characters. The possible values are:

- `IN GROUP`: Characters traveling together from the beginning.
- `JOIN GROUP`: Another character `[id` $\neq$ `0]` joins the primary traveler's group or the primary character `[id = 0]` midway through the journey.
- `EXIT GROUP`: Another character `[id` $\neq$ `0]` leaves the primary traveler's group or the primary character `[id = 0]` midway through the journey.
- `NO REL`: There is no travel relationship between the primary character `[id = 0]` and another character `[id` $\neq$ `0]`.